

**Special Section**

Mining the Metadata Quarries

New Metadata Standards for Digital Resources: MODS and METS

Metadata Generation: Processes, People and Tools

Data Collection for Controlled Vocabulary Interoperability – Dublin Core Audience Element

A Knowledge Network Constructed by Integrating Classification, Thesaurus and Metadata in a Digital Library

**Feature**

Spying and Secret Courts in America: New Rules and New Insights

# BULLETIN

*of the American Society for Information Science and Technology*

*December/January 2003*

Volume 29, No. 2

ISSN: 0095-4403 CODEN:BASICR

# Mining the

# METADATA

# QUARRIES



Thomas Edison,  
who invented the  
lightbulb, was  
afraid of the dark.

## Custom-Tailored Protection from ASIS&T: for What Keeps YOU Up at Night

**Y**ou're an individual, with your own hopes, dreams ... and worries: What happens if you become disabled and can't work for a year? Would your spouse be able to afford your family's current lifestyle if something happened to you? Will you be around to put your kids through college? If not, what then?

Unfortunately, it's not always possible to—like Thomas Edison—invent your own instant solutions to what keeps you up at night.

**ASIS&T can help.** As an ASIS&T member, you can choose from seven quality ASIS&T-sponsored insurance

plans to help safeguard you and your family.

Whether it's solid Term Life protection for you, Disability Income Insurance for you and your spouse, or Comprehensive HealthCare coverage for your whole family, the ASIS&T-sponsored Program is your one-stop source for custom-tailored protection for what matters most in **your life**. And each plan is available to members like you at **ASIS&T group-negotiated rates** that fit your budget.

Rest easy.... The ASIS&T-sponsored Insurance Program looks out for you.

Call 1-800-424-9883 for FREE information on the following ASIS&T-sponsored Plans:

- Term Life
- Medicare Supplements
- Disability Income
- Comprehensive HealthCare
- Member Assistance
- Catastrophe Major Medical
- High-Limit Accident

Sponsored by:



For FREE information about features, costs, eligibility, renewability, limitations and exclusions on any of the ASIS&T-sponsored Plans, call toll free 1-800-424-9883

The Term Life, Comprehensive HealthCare and High-Limit Accident Plans are underwritten by New York Life Insurance Company, 51 Madison Ave., New York, NY 10010. The Medicare Supplement Plans are underwritten by Monumental Life Insurance Company, Baltimore, MD. The Disability Income and Member Assistance Plans are underwritten by Unum Life Insurance Company of America. The Catastrophe Major Medical Plan is underwritten by The United States Life Insurance Company in the City of New York, Member American General Financial Group. All plans are administered by Marsh Affinity Group Services, a service of Seabury & Smith.

# BULLETIN

of the American Society for Information Science and Technology

## SPECIAL SECTION

### Mining the Metadata Quarries

11

#### Introduction

Stuart Sutton, Guest Editor

12

#### New Metadata Standards for Digital Resources: MODS and METS

Rebecca Guenther and Sally McCallum

16

#### Metadata Generation: Processes, People and Tools

Jane Greenberg

20

#### Data Collection for Controlled Vocabulary Interoperability – Dublin Core Audience Element

Joseph T. Tennis

24

#### A Knowledge Network Constructed by Integrating Classification, Thesaurus and Metadata in a Digital Library

Wang Jun

## Feature

8

#### Spying and Secret Courts in America: New Rules and New Insights

Lee S. Strickland

## IA Column

29

#### On Trust and Users

Andrew Dillon

## Departments

### 2 From the Editor's Desktop

Irene L. Travis

### 3 President's Page

Trudi Bellardo Hahn

### 4 Inside ASIST ASIST International Liaison

Julian Warner

#### Editor

Irene L. Travis

#### Publisher

Richard B. Hill

#### Advisory Board

Marjorie Hlava, chair;  
Irene Farkas-Conn; Sue O'Neill Johnson;  
Trudi Bellardo Hahn; Steve Hardin; Emil Levine;  
Kris Liberman; Lois Lunin; Ben-Ami Lipetz;  
Michel Menou; Linda Rudell-Betts; Candy Schwartz;  
Margarita Studemeister; Sheila Webber;  
Don Kraft, editor of *JASIST*, *ex officio*;  
Dick Hill, executive director of ASIST, *ex officio*.

#### ASIST Board of Directors

Trudi Bellardo Hahn, President  
Donald H. Kraft, Past President  
Samantha Hastings, President-elect  
Cecilia Preston, Treasurer  
Allison Brueckner  
Dudee Chiang  
Beverly Colby  
Andrew Dillon  
Abby Goodrum  
Douglas Kaylor  
Michael Leach  
Gretchen Whitney  
Karen Howell (Deputy)  
Richard B. Hill, Executive Director

The *Bulletin of the American Society for Information Science and Technology*, ISSN 0095-4403, is published bi-monthly, October through September, by the American Society for Information Science and Technology, 1320 Fenwick Lane, Suite 510, Silver Spring, MD 20910; 301/495-0900; Fax: 301/495-0810; e-mail: [asis@asis.org](mailto:asis@asis.org); <http://www.asis.org>

POSTMASTER: Send address changes to the *Bulletin of the American Society for Information Science and Technology*, 1320 Fenwick Lane, Suite 510, Silver Spring, MD 20910. The *Bulletin of the American Society for Information Science and Technology* is entered for periodicals postage paid at Silver Spring, MD, and at additional mailing offices.

The subscription rate for ASIST members is \$19, which is included in the annual membership dues. Non-member subscriptions, and additional member subscriptions, are \$60 per year U.S., Canada and Mexico; \$70 per year other, postpaid in the U.S. Single copies and back issues of the *Bulletin* may be purchased for \$10 each. Claims for undelivered copies must be made no later than three months following the month of publication.

Where necessary, permission is granted by the copyright owner for libraries and others registered with the Copyright Clearance Center (CCC) to photocopy any page herein for \$0.75 per page. Payments should be sent directly to CCC. Copying done for other than personal or internal reference use without the expressed written permission of the American Society for Information Science and Technology is prohibited. Serial-fee code: 0095-4403/83 \$0=\$0.75. Copyright © 2003 American Society for Information Science and Technology.

The American Society for Information Science and Technology (ASIST) is a non-profit professional association organized for scientific, literary and educational purposes and is dedicated to the creation, organization, dissemination and application of knowledge concerning information and its transfer.

The official ASIST journal is the *Journal of the American Society for Information Science and Technology*, published for the Society by John Wiley & Sons, Inc., 605 Third Avenue, New York, NY 10158.

The *Bulletin of the American Society for Information Science and Technology* is a news magazine concentrating on issues affecting the information field; management reports; opinion; and news of people and events in ASIST and the information community. Manuscripts are welcomed and reviewed for possible publication. Articles should not exceed 1500 words and may be accompanied by appropriate artwork. All manuscripts are subject to editing. Care is taken in the handling of manuscripts and artwork; however, the *Bulletin* cannot assume responsibility for lost or damaged material or for the return of unsolicited manuscripts. Send manuscripts to the *Bulletin of the American Society for Information Science and Technology*, 1320 Fenwick Lane, Suite 510, Silver Spring, MD 20910. Because the *Bulletin* is a news magazine, authors do not review galley proofs before publication.

Opinions expressed by contributors to ASIST publications do not necessarily reflect the official position of either their employers or the Society.



**Irene L. Travis**, Editor  
*Bulletin of the  
American Society for  
Information Science  
and Technology*  
Bulletin@asis.org

**T**his issue is one of two that will be devoted primarily to metadata and closely related topics, such as the Semantic Web. The second installment will appear in April/May 2003. I am very grateful to Stuart Sutton for editing the special section in this issue: **Mining the Metadata Quarries**.

I am struck by how sparsely even eight or ten articles cover this field now compared to five years ago when we published our first metadata issue (October/November 1997). Both the scope of the field and the number of projects have increased enormously, but some things don't change. We are still waiting to see any visible impact on the major Internet search sites from the hard labor in the metadata quarries. Jane Greenberg addresses one aspect of this problem when she looks at how metadata can be generated more efficiently and effectively and, in particular, at efforts to improve automatic metadata generation.

Another important thread in the current metadata world is XML. If you missed Jay ven Eman's excellent introduction to XML in our last issue you may want to backtrack.

These two themes, XML and metadata generation, come together in the article by Rebecca Guenther and Sally McCallum of the Library of Congress. Their article discusses MODS (Metadata Object and Description Schema), a development that provides a MARC-based XML schema that is less elaborate than the MARC 21 bibliographic format but more descriptive than the Dublin Core.

Another major theme in the metadata world is interoperability, which has many aspects. Stuart Sutton selected two of them for this special issue: value mapping and interoperability through schemas. Guenther and McCallum address this latter aspect in the second part of their article when they consider the METS proposal (Metadata Encoding and Transmission Standard), which provides a framework for attaching expanded types of metadata to digital resources, including administrative and structural information. Turning to research in value mapping, Joe Tennis tests the potential role of card sorting, using "audience" terms from seven different education vocabularies in his experiment.

Finally, Wang Jun explores ways to exploit library traditional tools, in this case the *Chinese Classification and Thesaurus*, in a networked environment. This article also concerns an XML experimental system.

Both Tennis and Wang are doctoral students, a category of potential author that we probably do not tap often enough. In addition, Wang is both the third place winner in the 2002 SIG/III International Paper Competition and a winner of one of the ASIST International Travel Grants for the Annual Meeting so generously provided this year by the Eugene Garfield Foundation. I hope that when you read this issue many of you will have had an opportunity to meet him and the other outstanding recipients in Philadelphia. As one of the judges I want to take this opportunity to express my deep gratitude, and I'm sure I speak for ASIST generally, to Sue O'Neill Johnson for chairing the Travel Grant Jury and easily doing enough work for four dedicated volunteers.

Which brings us to other matters international. In keeping with the efforts spearheaded so ably by SIG/III to increase our international presence, the Board has created an International Liaison position, whose first occupant is Julian Warner of Queen's College, Belfast. Julian and ASIST 2002 President Trudi Bellardo Hahn each provide details in this issue about his formidable assignment.

Nearer home, we are pleased to publish another insightful article from Lee Strickland, this one on the problem of balancing Fourth Amendment rights and homeland security. Lee focuses on the recent critical report by the Foreign Intelligence Surveillance Court to the United States Senate on the implementation of the Foreign Intelligence Surveillance Act and the USA Patriot Act.

Finally, Andy Dillon has some choice words to say about user-centered design in his IA column (I guarantee you won't want to miss this one!).

I'm glad to have seen so many of you at the ASIST Annual Meeting in Philadelphia and for those who could not attend, please look forward to full coverage of the meeting in the February/March issue of the *Bulletin*.



**Trudi Bellardo Hahn**  
2003 ASIST President  
User Education Services  
University of Maryland Libraries  
1103 McKeldin Library  
College Park, MD 20742  
301-405-9254  
th90@umail.umd.edu

It is a very good thing that a member of the Society, once elected president of ASIST, spends a year as president-elect. This is an astounding period of learning about all the behind-the-scenes work that goes on to sustain and grow the society. By “grow” I do not mean just increasing the membership (or at worst, keeping it from shrinking), although that is an important element. I mean mainly all the activities that keep ASIST alive, vibrant and at the leading edge of developments in the information field.

ASIST established task forces and planning groups this past year to consider ways to improve awards presentations at conferences, to construct the ASIST digital library, to update and implement the ASIST thesaurus, to review benefits to institutional members, to develop an applications magazine to supersede the *Bulletin* and to improve placement services, to name just a few. You can expect in the coming year to see the evidence of these efforts, and I hope that many of you will not only experience the benefits but will also get directly involved in expanding and sustaining them.

One new initiative that is just getting underway is in the area of international relations. During the past year a task force, composed of Steven Hardin (chair of Membership Committee), Michel Menou, Julian Warner, Executive Director Dick Hill and myself, proposed the establishment of a new position of International Liaison that reports to the Board of Directors. I have appointed the first liaison, Julian Warner. In Inside ASIST he shares his views and values about the liaison position and communication within the international information community.

The International Liaison has responsibilities for recruitment of international members to ASIST, retention of those members and assisting with and informing the Board about international issues. The following are examples of responsibilities of this position:

- Inform the Board of international developments and global information issues that represent opportunities or challenges for the Society, which might involve writing or commissioning briefing statements on issues that have implications for more than one jurisdiction.
- In cooperation with the Membership Committee, communicate with potential members and encourage them to join.
- Organize international receptions or other special events at conferences.

- Greet international attendees at conferences and help to make them welcome.
- Help international members to network with other ASIST members in their areas of interest (be a “linking agent”).
- Maintain contacts with international organizations on behalf of ASIST.
- Work closely with all SIGs in regard to their international information interests, but especially with SIG/International Information Issues, to ensure cooperation and collaboration as appropriate.
- Recruit other ASIST members to work with him/her in accomplishing these goals and objectives.
- Be highly visible, accessible and accommodating to others in ASIST who are interested in recruiting and supporting international members.

You will notice that the liaison does not work alone but is a bridge among several groups within ASIST, most notably the award-winning SIG/ International Information Issues and the Membership Committee, as well as headquarters staff. The liaison is also a communication channel to outside international organizations.

In selecting Julian Warner for this position, we looked for an ASIST member with deep international involvement who is willing to stay abreast of relevant international developments, who has excellent communication and interpersonal skills, tact and diplomacy and who can organize and manage a variety of assignments simultaneously. We expected that the individual would likely be an international member, but this was not required. Julian was selected as the liaison to launch the position and help the Board define its role.

Another exciting initiative will be a membership survey. In the past 20 years, the membership has been surveyed several times to gather varying amounts of data about members’ demographics, job and career interests and preferences for delivery of Society products – publications, conferences, continuing education and other professional benefits. However, no recent data has been collected, and we know that the membership profile and members’ interests have changed considerably in recent years. The current survey will be administered electronically, and we hope to get responses from a very high percentage of members. Planning future directions for the Society will depend very much on what you tell us – so please do respond!

# Inside ASIST

## Annual Meeting Coverage

The December/January issue of the *Bulletin of the American Society for Information Science and Technology* traditionally contains in-depth coverage of the ASIST Annual Meeting, including reports on major technical sessions, full photographic and textual coverage of the winners of the ASIST awards and news of other relevant activities.

This year, however, the mid-November date of the 2002 Annual Meeting fell outside the production cycle for this issue of the *Bulletin*. Therefore, the bulk of Annual Meeting News will be included in the February/March 2003 issue.

## 2002 ASIST Award Honorees

The following individuals and organizations were among those honored at the 2002 ASIST Annual Meeting Awards Banquet and at other venues during the meeting. Details of these honors will be included in the next issue of the *Bulletin*.

Award of Merit - **Karen Sparck Jones**  
Research Award - **Carol Tenopir**  
ISI Citation Research Award - **Chaomei Chen**  
Best Information Science Book Award - ***Beyond Our Control?* by Stuart Biegel**  
ISI Doctoral Dissertation Proposal Scholarship - **Joan Bartlett**  
UMI Doctoral Dissertation Award - **Pamela Savage-Knepshild**  
Pratt-Severn Award - **Elizabeth Zogby**  
James M. Cretsos Leadership Award - **Suzanne L. Allard**  
ISI Outstanding Information Science Teacher Award - **Deborah K. Barreau**  
Best *JASIST* Paper Award - **M. Kaszkiel and J. Zobel**  
Chapter Member-of-the-Year Award - **Stephen Spohn (Potomac Valley Chapter)**  
Chapter Electronic Publication-of-the-Year - **www.lacasis.org (Los Angeles Chapter)**  
Chapter Print Publication-of-the-Year - **OASIS (Los Angeles Chapter)**  
Chapter Event-of-the-Year - **Fall workshop of the Los Angeles Chapter**

## New Officers and Directors Join ASIST Board

With the start of the 2003 administrative year, the ASIST Board of Directors welcomed three new members and bade farewell to three whose terms have concluded.

Each of the new members will serve the society for three years. Those elected to the Board during the summer balloting are **Samantha Hastings**, president-elect; and **Allison Brueckner** and **Beverly Colby**, directors-at-large.

As the new members took their seats on the ASIST Board, **Trudi Bellardo Hahn**, elected last year as president-elect, assumed the presidency from **Donald H. Kraft**, who now serves as past president for one year.



**Hahn** is manager of Library User Education Services and adjunct professor in the College of Information Studies at the University of Maryland.

**Hastings** is an associate professor of information science and fellow of the Texas Center for Digital Knowledge at the University of North Texas in Denton. She teaches courses in digital image management, telecommunications, and science and technology reference. Her primary research interests



revolve around the problems of digital image storage and retrieval. Currently, she is the principal investigator for the Digital Image Managers Project, funded by a federal grant from IMLS.

An ASIST member since 1989, she has organized program sessions, chaired special interest groups (SIGs), served on committees and juries and on the Board of Directors as the SIG Cabinet Director for four years. She believes that the SIGs are the heart of ASIST and has worked diligently to help SIGs meet the challenges of working within a volunteer organization.

**Brueckner** received both her undergraduate and graduate degrees from Indiana University. She received her Masters in Information Science from Indiana University in December 1999. Since joining ASIST in 1997 as a student member from Indiana University, Allison has exhibited energy, dedication and leadership in several arenas. Allison has served as the



Indiana Chapter chair and program chair. She aggressively recruited quality speakers and presenters for chapter programs, initiated joint meetings with the Southern Ohio Chapter, updated and maintained the chapter website, began a review of the chapter bylaws and organized the first virtual meeting of a chapter executive committee. Since moving to Ann Arbor in 2000, Allison organized fellow ASIST members to restart the previously dormant Michigan Chapter, currently serving as chair. At the national level, Allison was a founding member and a driving force behind the creation of Special Interest Group/Digital Libraries, serving as the first chair. She set up the website for this new SIG and worked to get its listserv going. She has also served as a member of the Leadership Development Committee and is the chair of the Digital Libraries Task Force. She received the 2001 James M. Cretsos Leadership Award.

**Colby** is research consultant with the venture capital firm Flagship Ventures. Previously she was director of research with Battery Ventures. She holds the Master of Science degree in



Library and Information Science from Simmons College. Her ASIST participation includes both chapter and national activities. She was on the board of the New England Chapter for many years and, in the recent past, has worked with a committee to explore outreach for the chapter. Long active in SIG/Management, she currently serves as chair. She has served on a number of standing committees for ASIST, including Leadership Development, Awards & Honors and Nominations. She was one of the organizers of this year's Knowledge Management Summit which preceded the Annual Meeting.

As the new Board of Directors went to work in Philadelphia at the conclusion of the 2002 Annual Meeting, the service of those leaving the Board was acknowledged. **Joseph Busch**, immediate past president; and **Raya Fidel** and **Kris Liberman**, directors-at-large, were thanked for their distinguished efforts on behalf of the Society and its membership.

### News from ASIST Chapters

In October, the **Arizona Chapter** of ASIST was a co-sponsor of a colloquium entitled *Classification: The West and the Rest?* **Hope Olson**, professor, University of Alberta, presented findings from her research project on the cultural construction of classification. Over the summer, the chapter spearheaded a Virtual Community Building effort. The group experimented with various software to build virtual communities ranging from the simple to the complex; they had had fun identifying priorities and learning hands-on IT skills.

In September, the **Potomac Valley Chapter** of ASIST presented Liane Hansen, on-air host of NPR's Weekend Edition Sunday, in a session titled *Information, Accessibility and Terrorism's Impact on the Information Age*. Hansen has an extensive background in broadcast journalism, including more than 25 years of work as a radio producer, reporter and on-air host. She shared her views on information dissemination in today's environment, including the impact 9/11 has had on information accessibility and availability.

The **Northern Ohio ASIST Chapter (NORASIS)** joined with the Cleveland chapter of the Special Libraries Association to present the annual George Mandel lecture, this year featuring former ASIST president **José-Marie Griffiths**, speaking on *Leadership Skills for the Electronic Age*. The two orga-

### SOASIST Student Scholarship Essay Competition

Kitty McClanahan, a student in the College of Communications & Information Studies at the University of Kentucky, is the winner of the 2002 SOASIST Student Scholarship Essay Competition. The competition encourages students in Library and Information Science & Information Studies programs in Ohio and Kentucky to consider the benefits of participation in professional societies.

Competitors were asked to compose an essay addressing the following question: "What specific benefits do I expect to derive from attending the ASIST 2002 Annual Meeting?" Essays were judged by a panel of SOASIST members. The winner won reimbursement funding up to \$1000 for registration, airfare, food and lodging expenses to attend the 2002 ASIST Annual Meeting in Philadelphia.

This year's award was funded by LexisNexis.

Kitty's winning essay will be published in the upcoming issue of *soasis&t...on the move* ([www.asis.org/Chapters/soasis/enews/index.html](http://www.asis.org/Chapters/soasis/enews/index.html)).

nizations then cooperated on a session on *Careers in the Information Profession*, featuring a panel of speakers talking about their jobs in technical, business and non-traditional research fields and Marty Jaffe from Cuyahoga County Public Library's InfoPlace reviewing the latest resources and trends in job searching.

Meanwhile, the **Southern Ohio Chapter of the American Society for Information Science and Technology (SOASIST)** joined the Miami Valley Computing Societies to present their 14th Annual Fall Joint Meeting in September. *Disruptive Technology: A Panel Discussion on Innovation and Disorder* focused on technological changes that radically alter (i.e., "disrupt") business and societal paradigms.

SOASIST then teamed up with the **University of Kentucky Student Chapter of ASIST** to present Eric Weig, Digital Initiatives Librarian, University of Kentucky, on *Metadata and the Digital Library*. Weig, digital initiatives librarian at the University of Kentucky, is project manager for the Kentucky Virtual Library's Kentuckiana Digital Library Project.

The Seattle Reading Group of the **Pacific Northwest Chapter** started its third season in October with a discussion of the question, "What the heck are wikis and blogs?" The discussion articles for this session were "Blogs Take Web Diaries to the Next Level" ([www.cnn.com/2002/TECH/internet/05/09/blog/](http://www.cnn.com/2002/TECH/internet/05/09/blog/)) and "Build Community with Web Logs," ([www.clickz.com/tech/lead\\_edge/article.php/818321](http://www.clickz.com/tech/lead_edge/article.php/818321)).

Then in November, the Seattle Reading

Group discussed job hunting, focusing on "The Info Pro's Survival Guide to Job Hunting" and other articles in the July/August 2002 *Searcher* magazine (<http://www.infotoday.com/searcher/jul02/searcher>).

The **Los Angeles & New England Chapters of the American Society for Information Science and Technology** are offering the first ASIST chapter virtual program on distance education and e-learning. The free, online mini-course uses Blackboard (distance learning) software to discuss methods, technologies and issues of distance education and e-learning and features online video presentations by **Howard Besser**, UCLA, and **Richard Larson**, MIT. The program runs from December 9-20, 2002.

### News about ASIST Members

**Kevin Rioux**, formerly a visiting faculty member in the School of Library and Information Science at the University of South Florida, has joined the faculty of the Department of Library and Information Studies, School of Education, at the University of North Carolina at Greensboro. Rioux's research interests are information sharing in electronic environments, information acquiring behavior and human factors in interface design. He is currently a doctoral candidate at the University of Texas at Austin where he recently held the University Continuing Fellowship and the Robert R. Douglass Memorial Endowed Presidential Scholarship.

**Don Kraft**, immediate past president of ASIST and editor of the *Journal of the*

*American Society for Information Science and Technology*, was elected Fellow of the AAAS (American Association for the Advancement of Science) for distinguished contributions to information science and computer science through research and teaching and as editor of *JASIST*. This award is for those whose "efforts on behalf of the advancement of science or its applications are scientifically or socially distinguished."

**Eddie Rasmussen**, currently professor, School of Information Science, University of Pittsburgh, will be the Director of the

School of Library, Archival and Information Studies at the University of British Columbia, beginning on July 1, 2003. She was the program chair of the recently concluded ASIST Annual Meeting.

**Mary Lynn Rice-Lively** has been appointed to the newly created position of associate dean of the Graduate School of Library and Information Science (GSLIS) at the University of Texas at Austin. A former assistant to the Dallas City Manager and, more recently, assistant dean of information technology at GSLIS, Rice-Lively is an expert in

management and information technology and has held a variety of management positions during her 25 years of professional work.

**Deborah Barreau**, assistant professor at the School of Information and Library Science at the University of North Carolina at Chapel Hill, received the 2002 Steven I. Goldspiel Memorial Research Grant of the Special Libraries Association for her research proposal, "The New Information Professional: Vision and Practice." Barreau also received the 2002 ISI/ASIST Outstanding Information Science Teacher Award.

---

## OBITUARIES

### SAUL HERNER

Saul Herner, 79, a longtime ASIST member who did pioneering work in establishing chemical information systems, died of kidney failure October 11 at Friends Nursing Home in Sandy Spring. Herner retired in 1996 as president of Herner & Co., a consulting firm in Arlington, Virginia.

Herner, a Washington resident, was an early practitioner in the field of information science, which was fueled in large part by the interests of the federal government, researchers and the defense industry. He wrote about the subject in books and technical journals, defining it as the product of convergences of library science, computer and punch-card science, research and development documentation, abstracting, indexing and other fields and disciplines.

Herner, a native of Brooklyn, New York, moved to Washington in the early 1950s. The firm he started in the mid-1950s specialized in technical libraries, surveys and other information services for the government.

Mr. Herner earned a degree from the University of Wisconsin and took graduate courses in library science at New York University. He was a research chemist with the Army in the mid-1940s and then a chemical reference assistant at the New York Public Library. He was assistant curator and engineering librarian at New York University before moving to the Washington area to be chief librarian at the Applied Physics Laboratory of Johns Hopkins University. In the mid-1950s, he was head of the technical information and library planning group of Atlantic Research Co. Mr. Herner taught at Drexel Institute of Technology and the University of Maryland library school. He was editor in chief of a technical publishing firm, Information Resources Press and treasurer of Herner Analytics Laboratories in Rockville, which he co-founded with his brother in 1972.

His wife, Mary Alexander Herner, died in 1997. He is survived by his brother Albert E. Herner of Rockville.

### WINIFRED SEWELL

Longtime ASIST member Winifred Sewell died recently at her home in Maryland. She was 85. Sewell was senior librarian at Squibb Institute of Medical Research from 1946-1961. Subsequently, she went to Wellcome Research Labs at National Library of Medicine (NLM). She was instrumental in developing MEDLARS as Medical Subject Headings and later served as deputy chair of the Biological Services Division and head of the Drug Literature Program at NLM. Among many honors, she served as honorary president of the American Association of Colleges of Pharmacy in its 100th anniversary year, 1999-2000, the first woman and the first librarian to be so honored.

### ALBERT TABAH

Albert Tabah, 51, professor at the École de bibliothéconomie et des sciences de l'information (EBSI), passed away on September 30, 2002, following a long illness.

After a 15-year career in the medical, science and engineering libraries at McGill University and Concordia University, in 1990 Tabah returned to school in pursuit of a Ph.D. in information studies at McGill, a degree he earned in 1996. In 1993, he joined the faculty of EBSI and became assistant professor in 1996.

Tabah was a great success as a teacher and a scholar. He published notable research in the areas of bibliometrics in the sciences and in collection development. Both the World Bank and the International Development Research Centre sought his expertise, sending him on several missions to Senegal, Guinea and Burkina Faso.

A colleague, Marcel Lajeunesse, professor at EBSI, writes, "In the nine years he spent among us, Albert Tabah became friends with everyone who knew him. He was a highly valued colleague. We miss him."

**ASIST International Liaison**

by Julian Warner

**A**SIST has appointed an International Liaison. And as Kurt Gödel, known as a child as *Der Herr Warum*, might have asked, “Why?”

From a historical perspective, the intensification of international communication can be traced to the mid- and late-19th century developments in transport and communication. By the 1880s, it could be said that “the whole earth has been girded by telegraph cables” (by Marx in the third volume of *Das Kapital*). The late 19th century also marks the proliferation of international organizations and agreements, establishing reciprocity or commonality between national jurisdictions. The development of the world market and the Internet, now both painfully and pleasurably part of our everyday existence, can be traced to that period.

For information policies and activities, international communication presents certain issues. Most crucially, the ideas of jurisdiction and of national sovereignty have to be invoked and problematized. Justice Holmes noted in a seminal judgment of 1909

All legislation is *prima facie* territorial.... The general and almost universal rule is that the character of an act as lawful or unlawful must be determined wholly by the law of the country where the act is done.

How do we determine the location of communicative acts in cyberspace? What actions or sanctions can be taken against the producers of unlawful or unwanted communications, other than attempts at exile or banishment from the continuing conversation? Intellectual property can no longer be considered on a solely national basis, with international conventions going back to the 1880s, although with the much later assimilation of the United States.

What spheres of activity might fall within the remit of an international liaison, given this context? Addressing information policy issues which have transnational implications, liaising with other professional information societies, and within ASIST, complementing existing governance arrangements and establishing social connections with and between other international members would be significant.

An agenda that translates these concerns into specific actions must be guided by considerations of practicality and cooperation with existing arrangements. Otherwise, like the first exile, in one religious tradition, I may be forced to exclaim, “My punishment is greater than I can bear.” Initial steps would be writing articles for the *Bulletin of the American Society for Information Science and Technology* (as I’m doing here), recruiting an advisory group and developing a social presence at ASIST meetings. The overall aim of measures should be to facilitate the transformation of a community in itself to a community for itself, to assist the growth of the collective self-consciousness of the international information community. Could there be an analogy between the aim of developing an information community and the possible federal intention of developing the people of the United States of America into a collectivity fully known to itself by providing in the *Post Office Act of 1792* for the carriage of newspapers by the mail service “at extremely low rates”?

*Julian Warner is a faculty member in information science at the Queen’s University of Belfast, Northern Ireland, where he teaches courses in the human aspects of modern communication technologies and in information policy. Julian has been a member of ASIST since 1991. He is a member of SIG/International Information Issues, has served as chair of the European Chapter and of SIG/History and Foundations of Information Science and acted as a juror for the ISI Dissertation Scholarship Jury.*

**PUBLISHER’S STATEMENT**

United States Postal Service  
**Statement of Ownership, Management, and Circulation**

1. Publication Title <b>Bulletin of the American Society for Information Science and Technology</b>	2. Publication Number <b>0 0 9 5 - 4 4 0 3</b>	3. Filing Date <b>11/27/2002</b>
4. Issue Frequency <b>Bi-monthly</b>	5. Number of Issues Published Annually <b>6</b>	6. Annual Subscription Price <b>Members - \$19 Non-Members - \$60</b>
7. Complete Mailing Address of Known Office of Publication (Not printer) (Street, city, county, state, and ZIP+4) <b>1320 Fenwick Lane, #510, Silver Spring, MD 20910</b>		
8. Complete Mailing Address of Headquarters or General Business Office of Publisher (Not printer) <b>1320 Fenwick Lane, #510, Silver Spring, MD 20910</b>		
9. Full Names and Complete Mailing Addresses of Publisher, Editor, and Managing Editor (Do not leave blank) Publisher (Name and complete mailing address) <b>American Society for Information Science and Technology 1320 Fenwick Lane #510, Silver Spring, MD 20910</b> Editor (Name and complete mailing address) <b>Richard B. Hill</b> Managing Editor (Name and complete mailing address)		

10. Owner (Do not leave blank. If the publication is owned by a corporation, give the name and address of the corporation immediately followed by the names and addresses of all stockholders owning or holding 1 percent or more of total amount of stock. If owned by a partnership or other unincorporated firm, give its name and address as well as those of each individual owner. If the publication is published by a nonprofit organization, give its name and address.)

Full Name	Complete Mailing Address
<b>American Society for Information Science and Technology</b>	<b>1320 Fenwick Lane, #510 Silver Spring, MD 20910</b>

11. Known Bondholders, Mortgagees, and Other Security Holders Owning or Holding 1 Percent or More of Total Amount of Bonds, Mortgages, or Other Securities. If none, check box  None

Full Name	Complete Mailing Address
<b>None</b>	<b>n/a</b>

12. Tax Status (For completion by nonprofit organizations authorized to mail at nonprofit rates) (Check one)  
 The purpose, function, and nonprofit status of this organization and the exempt status for federal income tax purposes  
 Has Not Changed During Preceding 12 Months  
 Has Changed During Preceding 12 Months (Publisher must submit explanation of change with this statement)

PS Form 3526, October 1999 (Use Instructions on Reverse)

13. Publication Title <b>Bulletin of Amer. Society for Info.</b>	14. Issue Date for Circulation Data Below <b>Oct./Nov. 2002</b>
15. Extent and Nature of Circulation	
a. Total Number of Copies (Net press run)	
(1) Paid/Requested Circulation (Based on Form 3541, include advertiser's proof and exchange copies)	3600
(2) Paid In-County Subscriptions (Based on Form 3541, include advertiser's proof and exchange copies)	435
(3) Sales Through Dealers and Carriers, Street Vendors, Counter Sales, and Other Non-USPS Paid Distribution	2969
(4) Other Classes Mailed Through the USPS	0
c. Total Paid and/or Requested Circulation (Sum of 15a(1), (2), (3), and (4))	3404
d. Free Distribution Outside the Mail (Carriers or other means)	15
e. Total Free Distribution (Sum of 15d and 15e)	15
f. Total Distribution (Sum of 15c and 15f)	3419
g. Copies Not Distributed	147
h. Total (Sum of 15g and f.)	3566
i. Percent Paid and/or Requested Circulation (15c divided by 15g times 100)	99.56%
16. Publication of Statement of Ownership <input checked="" type="checkbox"/> Publication required: 1995 or printed in the USA <input type="checkbox"/> Publication not required	
17. Signature and Title of Editor, Business Manager, or Owner <i>Richard B. Hill</i> Publisher	Date <b>12/03/2002</b>

I certify that all information furnished on this form is true and complete. I understand that anyone who furnishes false or misleading information on this form or who omits material or information requested on the form may be subject to criminal sanctions (including fines and imprisonment) and/or civil sanctions (including civil penalties).

- Instructions to Publishers**
- Complete and file one copy of this form with your postmaster annually on or before October 1. Keep a copy of the completed form for your records.
  - In cases where the stockholder or security holder is a trustee, include in items 10 and 11 the name of the person or corporation for whom the trustee is acting. Also include the names and addresses of individuals who own or hold 1 percent or more of the total amount of bonds, mortgages, or other securities of the publishing organization. In item 11, if none, check box. Use blank sheets if more space is required.
  - Be sure to furnish all circulation information called for in item 15. Free circulation must be shown in items 15d, e, and f.
  - Item 15b. Copies not distributed, must include (1) newspaper copies originally stated on Form 3541, and returned to the publisher; (2) estimated returns from news agents; and (3) copies for office use, leftovers, spoiled, and all other copies not distributed.
  - If the publication had Periodicals authorization as a general or requester publication, this Statement of Ownership, Management, and Circulation must be published; it must be printed in any issue in October or, if the publication is not published during October, the first issue printed after October.
  - In item 16, indicate the date of the issue in which this Statement of Ownership will be published.
  - Item 17 must be signed.
- Failure to file or publish a statement of ownership may lead to suspension of Periodicals authorization.

PS Form 3526, October 1999 (Reverse)

# Spying and Secret Courts in America: New Rules and New Insights

by Lee S. Strickland

*Lee S. Strickland is a visiting professor in the College of Information Studies, University of Maryland; e-mail: leess@ucia.gov*

If there is any issue that galvanizes the American public in general as well as the academic and civil liberties communities, it is the role and judicial supervision of intelligence in American life. Indeed, it is a concern that spans the centuries – from our English ancestors to current citizens who experienced government excesses in the 1970s and remain concerned today, as government powers (authorities) have expanded in light of the war on terrorism with the USA Patriot Act of 2002.

Five hundred years ago English liberties were challenged by the court of the Star Chamber – an entity of the state that began benignly and publicly but evolved under the Stuart monarchs so that, by the reign of Charles I, its name had come to symbolize arbitrary, secretive proceedings crushing personal rights and liberty with no right of appeal. And although our Constitution contemplates that our judicial process today is fully open to the public, our Congress has created a secret U.S. Foreign Intelligence Surveillance Court (FISC) to mediate the tension between our individual rights and the President’s national defense authorities.

It was thus of great interest that on August 22, 2002, and for the first time, a lengthy opinion of the FISC in Washington was released to the public (see The FISC Opinion). Notably, it sheds detailed light on the practice of applying for Foreign Intelligence Surveillance Act (FISA) warrants, the sharing of the acquired intelligence information with criminal law enforcement authorities and the Administration’s latest arguments as to the scope of certain USA Patriot Act authorities. What we see from the insights in this opinion are information management difficulties in past cases that have raised significant judicial concerns as well as a riveting political and legal debate today concerning

the use of intelligence authorities by law enforcement. What we will decide as a nation is, quite likely, nothing less than the proper scope of intelligence in our democratic society.

## The FISA Process and Problems in Brief

Readers will recall our previous *Bulletin* article (February/March 2002) that discussed the FISA and the changes effected by the recent USA Patriot Act. In brief, the FISA regulates the collection of “foreign intelligence information” (see Definition of Foreign Intelligence Information) from foreign powers or agents of foreign powers through a complex scheme of Attorney General (AG) approvals and in most cases applications to the FISC for secret warrants. The threshold condition for such warrants is not probable cause of criminal activity as with traditional Fourth Amendment law, but rather probable cause that the target is a foreign power or an agent of a foreign power. Thus, in light of the fact that FISA warrants can authorize electronic surveillance and physical searches and may be

## The FISC Opinion

The opinion was released by the Senate Judiciary Committee having received it from the FISC in response to a committee inquiry dated July 31, 2002. The committee has been examining how the FISA is working in practice and specifically considering the Department of Justice requests for investigative powers beyond those granted by the USA Patriot Act. A copy of the opinion is available at [www.washingtonpost.com/wp-srv/onpolitics/transcripts/fisa\\_opinion.pdf](http://www.washingtonpost.com/wp-srv/onpolitics/transcripts/fisa_opinion.pdf) as well as [www.fas.org/irp/agency/doj/fisa/fisc051702.html](http://www.fas.org/irp/agency/doj/fisa/fisc051702.html).

### The Definition of Foreign Intelligence Information

The FISA defines *foreign intelligence information* as information about (1) an actual or potential attack or other grave hostile acts of a foreign power, (2) sabotage or international terrorism by a foreign power or an agent of a foreign power, (3) clandestine intelligence activities by a foreign power or agent or (4) a foreign country that is necessary to the national defense or the security of the United States or the conduct of the foreign affairs of the United States.

directed against both aliens and U.S. persons, the constitutional concern has always been that this process should not erode our basic rights under the Fourth Amendment.

It is this concern that led to the minimization requirements in the FISA and implementing AG procedures – to constrain the acquisition and retention and prohibit the dissemination of non-publicly available information concerning U.S. citizens or permanent resident aliens if it does not concern “foreign intelligence.” These procedures, however, specifically allow for the retention and dissemination of information that relates to terrorism or is evidence of any other crime that has been, is being or is about to be committed. Thus, this is the limited authority to share intelligence-acquired information in the United States with law enforcement authorities but also restriction that has heretofore prevented law enforcement authorities from having direct access to FISA files and from directing or otherwise controlling FISA investigations toward law enforcement objectives. If this were not the case, the easier-to-grant FISA warrants could become the norm for any criminal case having some “foreign intelligence” aspect.

In recognition of the important constitutional considerations at issue, the FISC has established a “wall” procedure for all FISA orders that requires (a) certification that the purpose is foreign intelligence, (b) disclosure of all criminal information aspects of an intelligence case including specifics on information sharing with law enforcement and (c) designation of a senior official to moderate the flow of information to law enforcement including the FISC itself in significant overlapping criminal and intelligence cases. Again, the purpose was “to preserve both the appearance and the fact that FISA surveillance and searches were not being used sub rosa for criminal investigations” and to prevent prosecutors from becoming “de facto partners in FISA searches.” [FISC Order of May 17, 2002]. However, while these constitutional limitations and processes appear to many as clear and reasonable, the FISC over the last two years has had serious concerns about inaccurate FBI affidavits in FISA applications – errors that have misled the Court as to actual law enforcement purposes and information sharing in at least 75 FISA cases. And, although the FISC directed the Department of Justice to conduct an investigation, no report was forthcoming.

### The USA Patriot Act Changes and Attorney General’s Implementations

Among the many changes effected by the USA Patriot Act and particularly relevant here, the basis for granting FISA orders was changed from foreign intelligence being the “primary” purpose to a “significant” purpose. Why this seemingly technical change? It reflects the reality that most terrorism investigations have both intelligence and law enforcement aspects and Congress wanted to ensure that the existence of a law enforcement proceeding would not legally imperil a continuing intelligence investigation. The Congress also took the related step of ensuring that information could be effectively shared between intelligence and law enforcement personnel. While the law had not really prevented such exchanges, a perception had grown among officials, as evidenced by recent statements and testimony by FBI officials, that there were significant limitations. Thus, the USA Patriot Act expressly authorized intelligence officers who are using FISA *to consult* with federal law enforcement officers and *to exchange a “full range of information and advice.”* [See 50 U.S.C. §§ 1806(k) and 1825(k).] Together, the intent of these provisions was to reduce the conundrum between intelligence and law enforcement personnel as to the path of an investigation and to make clear that a FISA order could be obtained even if a criminal prosecution was contemplated or indeed had been instituted during the pendency of an intelligence investigation.

Few thought at the time of passage that these provisions would constitute a new grant of positive authority to law enforcement personnel – but they did and led to the FISC order that is the subject of this article. Notwithstanding the concerns of the FISA Court as to inaccurate FBI submissions, the AG proposed on March 6, 2002, new FISA minimization procedures including giving criminal prosecutors access to *all information developed* in FISA cases and allowing criminal prosecutors to consult and provide advice and recommendations to intelligence officials regarding any intelligence case including the *initiation, operation, continuation or expansion of FISA searches.* The proposed March 2002 procedures are available at [www.fas.org/irp/agency/doj/fisa/ag030602.html](http://www.fas.org/irp/agency/doj/fisa/ag030602.html).

The *increased sharing*, as we have seen, is certainly authorized by the terms of the USA Patriot Act. The *direct use of the FISA by law enforcement* is, however, more problematic and is supported only by the Administration’s argument predicated on the “significant purpose” change – specifically, that the change contemplates dual intelligence and law enforcement objectives and hence dual use. Accordingly, a foreign counter-terrorism (CT) intelligence purpose can be met by the fact of a criminal prosecution for terrorism and hence such a criminal prosecution may make direct use of the FISA. The contrary argument is that the dual objective is correct but that FISA authorizes use only by intelligence officials and that use by criminal authorities eviscerates our Fourth Amendment rights requiring warrants based on probable cause of criminal activity.

In considering the proposed direction, the FISC approved the sharing provisions as provided by the USA Patriot Act but held that the direct use of FISA authorities by law enforcement officials represented a substantial and constitutional issue. Specifically, the court held that the AG directive would give criminal prosecutors a significant role in “directing” FISA investigations, would violate the statutory basis for FISA orders (intelligence purpose) and would effectively substitute the FISA for the Fourth Amendment. Accordingly, the FISC revised the proposed AG procedures in part to allow the new levels of coordination and sharing but to retain the critical aspects of the “wall” that precludes direct use or control of FISA authorities by law enforcement personnel. Hence, to be maintained for the moment is the “bright line” between criminal and intelligence investigations and the rights of U.S. persons under these two bodies of law.

### What Do These Developments Mean for Us as Citizens and Information Professionals?

First, for the moment, we preserve Fourth Amendment values while still allowing the most effective intelligence and law enforcement operations. In no manner does the FISC order compel the United States to be impotent in the face of terrorism. Intelligence and law enforcement investigations of the same target can proceed in tandem with full coordination and sharing. If and when critical information is developed – either through intelligence or law enforcement efforts – the United States can employ a panoply of tools to act against the danger presented. If there is sufficient evidence of terrorism or other U.S. crime, a prosecution can ensue and, if there is not, deportations and/or foreign prosecutions are all available options. These limitations simply mean that criminal prosecutors may not freely use the FISA process in lieu of established constitutional search and seizure process.

Second, we must be prepared to join in the continuing legal and political debate over the proposed procedures. On August 23, 2002, the Department of Justice appealed this decision to the special three-judge court that was created to review FISC decisions. This court – the U.S. Foreign Intelligence Surveillance Court of Review – has not met previously and consists of three federal appellate court judges appointed on a rotating basis by the Chief Justice. A decision is pending at the present time and further review is also available by the Supreme Court. And, irrespective of the final judicial decision, there will be further political action. In one of the more unusual aspects of this controversy, the Government brief on appeal (available at [www.fas.org/irp/agency/doj/fisa/082102appeal.html](http://www.fas.org/irp/agency/doj/fisa/082102appeal.html)) asserted that Senator Leahy, identified as the “drafter of the coordination amendment” in the Patriot Act, “had agreed” that there is no longer a distinction between using FISA for a criminal prosecution and using it for foreign intelligence collection. But in a statement on September 10, 2002, in the context of a Senate Judiciary Committee hearing, the Senator clearly renounced the assertion that he supports the Attorney General’s position:

That was not and is not my belief. We sought to amend FISA to make it a better foreign intelligence tool. But it was not the intent of these amendments to fundamentally change FISA from a foreign intelligence tool into a criminal law enforcement tool. We all wanted to improve coordination between the criminal prosecutors and intelligence officers, but we did not intend to obliterate the distinction between the two, and we did not do so. Indeed, to make such a sweeping change in FISA would have required changes in far more parts of the statute than were affected by the USA Patriot Act . . . [and] such changes would present serious constitutional concerns.

And, perhaps beyond the legal debate, another question is presented – the practical necessity, not to mention the political wisdom, of the proposal. It is not at all clear why full sharing and coordination is insufficient, why this authority is truly needed and why we should as a nation detract our attention from the real challenge of terrorism with a perceived threat to civil liberties. In sum, perhaps the proposal harms U.S. interests by eroding public support for the counter-terrorism mission.

Third, we must plan in our businesses for the possible receipt of FISA orders since this debate does not affect the basic availability and terms of such orders. Information-centric institutions – schools, libraries and Internet service providers – will continue to receive secret FISA warrants and recipients will be barred from disclosing the terms. Of course, this does not mean that recipients are prevented from fully disclosing the fact and content to management, directors and legal counsel for their institutions. A more open question is whether recipients may disclose the simple fact that their institutions have received one or more FISA warrants; authorities are mixed but the weight tends to support the right to make this very limited statement.

And fourth, we should consider whether information exploitation tools – more than problematic legal arguments – could best facilitate the effectiveness of dual law enforcement and intelligence cases. For example, would visualization tools allow us not only to understand the data better but also track the originating interests and the using interests? Would time series analysis tools allow much the same tracking but include the ability to understand how information builds a case? And lastly, would not commercially available collaborative tools provide a mechanism to share and also track information relevant to multiple cases? Such tools could provide an operational benefit not only to concerned government officers but also to the FISC and government managers to ensure their appropriate oversight and adherence to minimization requirements. In sum, many of the errors of the past that greatly concern the FISC and limit investigative efforts today, might be ameliorated by the judicious application of information technology.

# Special Section

## Mining the Metadata Quarries

by Stuart A. Sutton, Guest Editor

---

*Stuart Sutton is an associate professor in the Information School of the University of Washington. He is co-chair of the Dublin Core Metadata Initiative Education Working Group and serves on the initiative's advisory and usage boards. He can be reached at [sasutton@u.washington.edu](mailto:sasutton@u.washington.edu).*

---

It is a foregone conclusion that metadata describing digital resources in a globally networked environment will be a bedrock for the Semantic Web. The four articles in the special section of this issue of the *Bulletin* touch on a few of the threads in metadata research, development and deployment. While for many the notion of metadata is a simple one and hardly new, the proliferation in a globally networked environment of metadata schemas expressing the needs of discourse and practice communities as well as organizations throughout the private and public sectors taxes our understandings of semantic interoperability. Research and development cover many different areas including interoperability among complex schema and value spaces; application profiles that permit the wedding of metadata statements from disparate schemas through network-accessible schema registry services; and metadata support tools. In this issue we look at a sample of current work.

Rebecca Guenther and Sally McCallum examine the rationale and the basic architecture of the Metadata Object and Description Schema (MODS) and the Metadata Encoding and Transmission Standard (METS). MODS is intended as a bridge between the descriptively rich MARC 21 metadata and terse metadata schema such as the Dublin Core Metadata Element Set. The authors describe the basic features of MODS, its prospective users and current experimentation in its use. METS is an XML-based schema that provides a means for packaging, or pointing to, descriptive, administrative, structural, rights and behavioral metadata for digital resources. METS supports the seamless flow

of metadata and electronic resources between networked systems.

Jane Greenberg's article defines a framework for identifying and examining processes of metadata generation, the tools required for that generation and the categories of metadata creators – human and automatic. In addition to providing class definitions for the various entities in her metadata generation framework, Greenberg provides references to examples of metadata generation tools and research and development projects. The article includes a brief description of the Metadata Generation Research Project at the University of North Carolina studying the entities in the framework and developing protocols for collaborations in its various processes.

The final two articles focus on value space problems. Joseph Tennis describes a pilot study that explores the interoperability problem faced by multiple metadata projects making statements by means of disparate controlled vocabularies. The goal of the study is to test a methodology for developing a switching language using card sorts, cluster analysis and talk-aloud protocols. Wang Jun's article discusses the tools, architecture and the implementation of an experimental system to support knowledge management. He describes development of a concept network consisting of nodes and edges developed from the *Chinese Classification and Thesaurus*, an integrated product widely used in Chinese libraries, to which metadata records are bound. Research includes expanding and customizing the concept network using information from the titles from the bibliographic records. The test database is in computer science.

# New Metadata Standards for Digital Resources: MODS and METS

by Rebecca Guenther and Sally McCallum

*Rebecca Guenther and Sally McCallum are with the Network Development and MARC Standards Office at the Library of Congress. Rebecca Guenther can be reached at [rgue@loc.gov](mailto:rgue@loc.gov)*

**M**etadata has taken on a new look with the advent of XML and digital resources. XML provides a new versatile *structure* for tagging and packaging metadata as the rapid proliferation of digital resources demands both rapidly produced descriptive data and the encoding of more *types* of metadata. Two emerging standards are attempting to harness these developments for library needs. The first is the Metadata Object and Description Schema (MODS), a MARC-compatible XML schema for encoding descriptive data. The second standard is the Metadata Encoding and Transmission Standard (METS), a highly flexible XML schema for packaging the descriptive metadata and various other important types of metadata needed to assure the use and preservation of digital resources.

## MODS Development

The Library of Congress' Network Development and MARC Standards Office developed MODS ([www.loc.gov/standards/mods/](http://www.loc.gov/standards/mods/)) in consultation with interested experts to satisfy the expressed need for an abbreviated XML version of MARC 21. XML is being increasingly deployed in computer applications, particularly on the Web, as a richer, more flexible alternative to HTML. Many have expressed the need to move to XML for metadata in libraries and other cultural institutions. It is appropriate for an XML version of MARC to be investigated since it is perhaps the oldest metadata standard designed for use in computers.

Over the years people have expressed concerns about the number of data elements in MARC and their complexity. Some have suggested use of the Dublin Core Metadata Element Set ([\[core.org\]\(http://dublincore.org\)\), although that set is intended to satisfy a broader range of purposes and communities than MARC 21. In order to address these concerns about MARC and also allow for a rich description, the Library of Congress developed MODS, an XML schema with language-based tags that includes a subset of data elements derived from MARC 21. It is intended to carry selected data from existing MARC 21 records as well as to enable the creation of original resource description records.](http://dublin</a></p></div><div data-bbox=)

## Features

MODS is intended to complement other metadata formats and to provide an alternative between a simple metadata format with a minimum of fields and little or no substructure such as Dublin Core and a very detailed format with many data elements having various structural complexities such as MARC 21. MODS has a high level of compatibility with MARC records because it inherits the semantics of the equivalent data elements in the MARC 21 bibliographic format. Thus, it is richer than Dublin Core and more compatible with library data than ONIX ([www.editeur.org/onix.html](http://www.editeur.org/onix.html)), which was developed for the book industry, but it is also simpler than the full MARC format (either as ISO 2709 or full MARCXML). It is more "friendly" because it uses language-based tags that can be easily understood by anyone dealing with the "raw" record, as opposed to the numeric tags traditional to MARC.

Most elements that have been defined in MODS have equivalents in the MARC 21 bibliographic format. In addition the Library of Congress has made available mappings between MARC and

MODS and vice versa ([www.loc.gov/standards/mods/mods-mapping.html](http://www.loc.gov/standards/mods/mods-mapping.html); [www.loc.gov/standards/mods/mods2marc-mapping.html](http://www.loc.gov/standards/mods/mods2marc-mapping.html)). Since MODS elements inherit the semantics of MARC elements, an element in MODS has the meaning detailed in the MARC 21 bibliographic format.

In MODS some elements in MARC have been repackaged, for example in cases where several data elements are brought into one. This repackaging occurs in the MODS element *genre*, which uses controlled values that are used in various MARC elements, particularly in fixed fields. The Library of Congress has made available a controlled list of genre values found in various places in the MARC 21 bibliographic format to be used with the MODS *genre* element ([www.loc.gov/marc/sourcecode/genre/genrelist.html](http://www.loc.gov/marc/sourcecode/genre/genrelist.html)).

MODS, like MARC 21, does not assume the use of any particular cataloging code. It can accommodate record content that is full AACR2 with authoritative name and subject headings, uncontrolled by cataloging rules, or anything in-between.

Since MODS is a subset of MARC, decisions were made about which elements to include, which to combine with others to form a single element and which to drop altogether. For instance, there are numerous types of relationships that are expressed in the MARC linking entry fields. These are carried in MODS under *relatedItem* with a type attribute to express the type of relationship. Not all relationships in MARC are given type values.

Certain MODS elements define concepts that recur in more than one element as sub-elements. XML facilitates using the same definition for multiple elements. For example, “name” can be the primary name associated with the resource or a name associated with a related item; in MODS, both use the same definition. This concept is certainly present in MARC 21 but not as consistently as in MODS.

Since MODS includes a subset of MARC 21 bibliographic fields, it allows for a conversion from MARC 21 fields to MODS, while other MARC 21 fields may be dropped or carried in a less specific manner. The MODS schema does not target “round-tripability” with MARC 21. A converted record may lose some of its tagging, for instance, when the tagging is simpler, or accommodate some data even when there is not an equivalent data element. When an XML schema is desired that does not result in any data loss, the MARC 21 XML schema may be used ([www.loc.gov/standards/marcxml/](http://www.loc.gov/standards/marcxml/)), since it allows for the expression of a full MARC record in XML. For any conversion between MARC in ISO 2709 format and MODS, it is expected that the record would first go through a conversion to MARCXML before a transformation to the subset that is MODS. The Library of Congress is pro-

viding tools for the conversion from MARC 21 to MARCXML with a further transformation to MODS.

### Prospective Uses

The need for a rich metadata standard such as MODS has been expressed by members of the digital library and related communities as they attempt to implement projects involving search and retrieval, management of complex digital objects, integrating metadata from library databases with other non-MARC sources and other functions.

The “Search/Retrieve Web Service” (SRW) in ZING (Z39.50 International Next Generation) ([www.loc.gov/z3950/agency/zing/srwu/srw.html](http://www.loc.gov/z3950/agency/zing/srwu/srw.html)) is a proof-of-concept initiative to develop value-added search and retrieve applications built on Z39.50 along with Web technologies – XML, SOAP/RPC and HTTP. It defines a search service that specifies metadata

schemas for retrieval. Since it uses XML, an XML metadata schema is needed, and one compatible with library data such as MODS would be desirable.

The Open Archives Initiative Protocol for

Metadata Harvesting ([www.openarchives.org/](http://www.openarchives.org/)) harvests MARC records from multiple systems and makes them available widely. Generally, the records have been available in MARC (using MARC tagging and syntax in MARCXML) or simple Dublin Core in XML. The Library of Congress is planning to incorporate MODS as an alternate format for its over 100,000 metadata records that describe various forms of material digitized for American Memory. This will allow for the export of richer metadata than the Dublin Core record, which drops much of the metadata, but provides simpler data than full MARCXML.

MODS may be used for original resource description that allows for rich description that is generally compatible with existing library data and is expressed in XML syntax. Because it includes a subset of MARC fields and repackages some of them, it is particularly useful for technician input.

An additional use of MODS is as an extension schema for descriptive metadata for a METS object, as detailed later in this paper.

### Experimentation with MODS

Since MODS was officially made available in June 2002, experimentation is just beginning. In June 2002, MODS was frozen for a six-month trial, although suggested additions are being listed on the MODS website. The following describes a few sample experiments.

- *The Library of Congress’ Audio-Visual Prototyping Project* is exploring aspects of digital preservation for audio and

## MODS: Metadata Object and Description Schema, a MARC-compatible XML schema for encoding descriptive data.

video. This collaborative project is developing approaches for packaging digital content, with a focus on metadata (<http://lcweb.loc.gov/rr/mopic/avprot/avprhome.html>). The project is experimenting with METS for packaging the digital object and its metadata and is implementing MODS for use as its descriptive metadata schema, particularly because of the rich descriptions of relationships with other items that may be expressed. Where possible metadata is being reused from descriptive cataloging records in one of the Library's databases with minimal data loss. In some cases, original resource description is provided and a MODS template is used.

- *MINERVA* (*Mapping the Internet: Electronic Resources Virtual Archive*) is an experimental pilot developed to identify, select, collect and preserve websites (<http://lcweb.loc.gov/minerva/minerva.html>). LC is collaborating with the Internet Archive (Alexa), SUNY and the University of Washington to collect and archive websites, providing descriptive metadata that will be used to search, retrieve and analyze the archived collections. Metadata will be created for websites in the collection using the MODS schema because of its compatibility with MARC data, to be used in the search and retrieval system and later converted to MARC and added to the Library's online catalog.
- The University of Chicago Press is implementing a project to support the development of the *Chicago Digital Distribution Center*, which would be built upon its traditional distribution center and involves making digital books available for distribution. The Press harvests MARC records to enhance searchability and for export to their client presses, converting them into MODS for more concise description and more understandable language-based tags.
- The *California Digital Library* is establishing a generic METS repository infrastructure to help manage the digital objects in its control. A project provides for search and display of 1,500 records for books published online by the CDL on behalf of the University of California Press. Records are extracted from the union catalog and transformed to MODS, then inserted into the METS record. Specified fields are used for indexing and searching as well as in response to an Open Archives Initiative Harvesting Protocol query.

## METS: Metadata Encoding and Transmission Standard, a highly flexible XML schema for packaging the descriptive metadata and various other important types of metadata needed to assure the use and preservation of digital resources.

### METS

The Metadata Encoding and Transmission Standard (METS) ([www.loc.gov/standards/mets](http://www.loc.gov/standards/mets)) grew out of several experimental 1990s digital projects. In February 2001, the Digital Library Federation convened a meeting of experts from several projects to evaluate what had been learned with respect to metadata and to decide how to go forward. Out of that meeting came the idea for METS, an XML document that packages the metadata associated with a digital resource – the descriptive, administrative, structural, rights and other data needed for retrieving, preserving and serving up digital resources. Then, in a little over a year the METS XML schema was developed, a maintenance structure set up and experimentation worldwide began.

METS metadata is essential for a digital material repository, where digital resources – over 7 million at LC alone – are stored along with information about the resources. A repository, which can take many configurations, is the instrument for access and preservation of the objects. The METS data is also important for the interchange of digital objects for viewing and use by other systems. If the digital resource has with it the METS description, the file should be usable for many activities at the receiving system.

### Characteristics of METS

METS is an open standard, not a proprietary one. Library system staff and librarians who also participate in developments in the Internet community are constructing it. Jerome McDonough of NYU serves in the critical role of editor-in-chief of the METS XML schema. The Library of Congress has agreed to serve as the maintenance agency for the standard — building a website for METS, supporting an open list-serv for implementers and working on extension schema. Recently an editorial board was formed of major contributors to the development thus far with the intent to identify and bring in global partners. The schema is now complete and stable enough to consider taking it to a formal standards body such as ISO or NISO.

The structure of the METS schema is highly flexible and relatively simple. It is conceptually six modules that contain and/or point to the different types of metadata needed for a digital resource. In several of the modules the METS standard does not define the metadata elements and tags to be used. It allows the user to choose a standard “extension”

schema and identify and use it. For several modules it allows the metadata to reside outside the package, pointed to from within the METS document. These features exemplify the flexibility of METS.

### The METS “Package”

The six parts of the METS package or document are as follows: header, descriptive metadata, administrative metadata, file section, structural map and behavior section. They are all optional except for the header and structural map, which are needed for basic access to the digital resource. The descriptive, administrative and behavior sections may reside in the METS document or be external. If they are internal, XML schemas are preferred. If the metadata for these sections is external and merely pointed to from the appropriate section of the METS document, the metadata may be of any type and format. For descriptive metadata it may even be an entry in a catalog if the catalog record can be adequately referenced.

The METS descriptive metadata section is the most familiar to librarians, as it contains cataloging and finding-aid data. There are several established schemas that can be used for the descriptive metadata – including MODS, which was designed with a special focus on electronic resources; Dublin Core, when only minimal data is needed; or MARCXML when full MARC record information is available.

The administrative metadata is the most critical for use and preservation of the digital resource. Here resides source information such as resource creation date, resource format information, resource use information, digital provenance and copyright and license information. This section may contain information on past transformations and migrations of the data and master/derivative information, all useful for preservation purposes. XML schemas are not yet standard for most of this information but the METS project is pushing their development. For example, the recently completed NISO data dictionary for technical metadata for still images ([www.niso.org/standards/resources/Z39\\_87\\_trial\\_use.pdf](http://www.niso.org/standards/resources/Z39_87_trial_use.pdf)) is being used through a supporting XML schema, MIX (Metadata Imaging in XML) ([www.loc.gov/mix/](http://www.loc.gov/mix/)). Project participants also have drafts prepared for the technical data for text, audio and moving images. The draft schemas are available from the METS website.

The behavior section of the METS document contains pointers to computer programs or applications that are used to display digital objects such as page-turners or audio players. The behavior information is intended to assist in providing “disseminators” for end user access.

The header, file group and structural map sections may only reside in the METS document – there is no option to point to them outside. The file section identifies all the files the object clusters, such as thumbnails, master archival, pdf versions and text-encoded versions. The structural map contains a clear layout of the hierarchical structure of the document. An important feature of METS is that the structural

map may point to parts of the descriptive and administrative metadata from different places on the structure hierarchy of the resource, enabling linking to subparts of digital resources, such as cuts on sound recordings. The header also gives information about the METS document itself, such as identifiers, date created, dates updated and status.

METS allows use of any established schemas for the different modules. While essential for acceptance in these times when standards are not yet in place, such flexibility could impede interoperability. Accordingly, the library community will be trying to work out a subset of schemas or profiles for the different types of data described above that will be used across exchange communities. This makes international participation in the use and development of METS especially important.

### METS Implementation and Use

METS came at a key moment for implementation and use. Many institutions had experimented with digitization and had begun to build collections large enough to seriously require better organization tools. Also there has recently been an increasing number of projects for archiving of open access Web subsets. The recent METS implementations have fortunately varied in form of material and size, giving good information to the standard’s developers. To indicate only a few projects, the Library of Congress is using METS for a very large body of moving image and audio material and other mixed media folk life resources; the National Library of Wales is using METS initially for textual material; Harvard is experimenting with audio collections; and Michigan State is working with moving images. Both OCLC and RLG are working METS into their digital projects.

### Conclusions

The library community has well developed bibliographic description traditions that with some adjustment for digital resources, such as the MODS development exemplifies, will serve the digital future. In the larger metadata picture, the development of METS is a big step toward bringing to non-descriptive metadata the stability needed for a smoothly functioning Internet environment where electronic resources flow seamlessly between systems. These developments relate well to the OAIS (Open Archival Information System) Reference Model ([www.ccsds.org/RP9905/RP9905.html](http://www.ccsds.org/RP9905/RP9905.html)), which helps to define the processes and boundaries in creating, managing, sustaining and serving digital resources. The METS package can be used to collect digital resource metadata for submission to the repository, serve as the place for the metadata within the repository and be the supplier of information to the tools that provide the resources to the patrons.

---

This article is based on one to appear in early 2003 in *Portal: Libraries and the Academy*, Johns Hopkins University Press.

# Metadata Generation: Processes, People and Tools

by Jane Greenberg

*Jane Greenberg is assistant professor, School of Information and Library Science, University of North Carolina, Chapel Hill. She can be reached by e-mail at [janeg@ils.unc.edu](mailto:janeg@ils.unc.edu)*

**M**etadata generation is the act of creating or producing metadata. Generating good quality metadata in an efficient manner is essential for organizing and making accessible the growing number of rich resources available on the Web. The success of digital libraries, the sustenance of interoperability – as promoted by the Open Archives Initiative – and the evolution of Semantic Web *all* rely on efficient metadata generation. This article sketches a metadata generation framework that involves *processes, people* and *tools*. It also presents selected research initiatives and highlights the goals of the *Metadata Generation Research Project*.

## Metadata Generation Processes

In today's networked environment metadata is produced by both *human* and *automatic* processes. **Human metadata generation** takes place when a person such as a professional metadata creator or content provider produces metadata. The quality of human generated metadata is often determined by semantic and syntactic adherence to a metadata schema specification. Historically the only form of metadata creation, this process still dominates libraries, museums, archives and other information resource centers. Human metadata generation is popular on the Web as demonstrated by a notable increase in the supply of "keyword" and "description" XHTML META tags. Further evidence is Adobe's recent enhancement of XMP (eXtensible

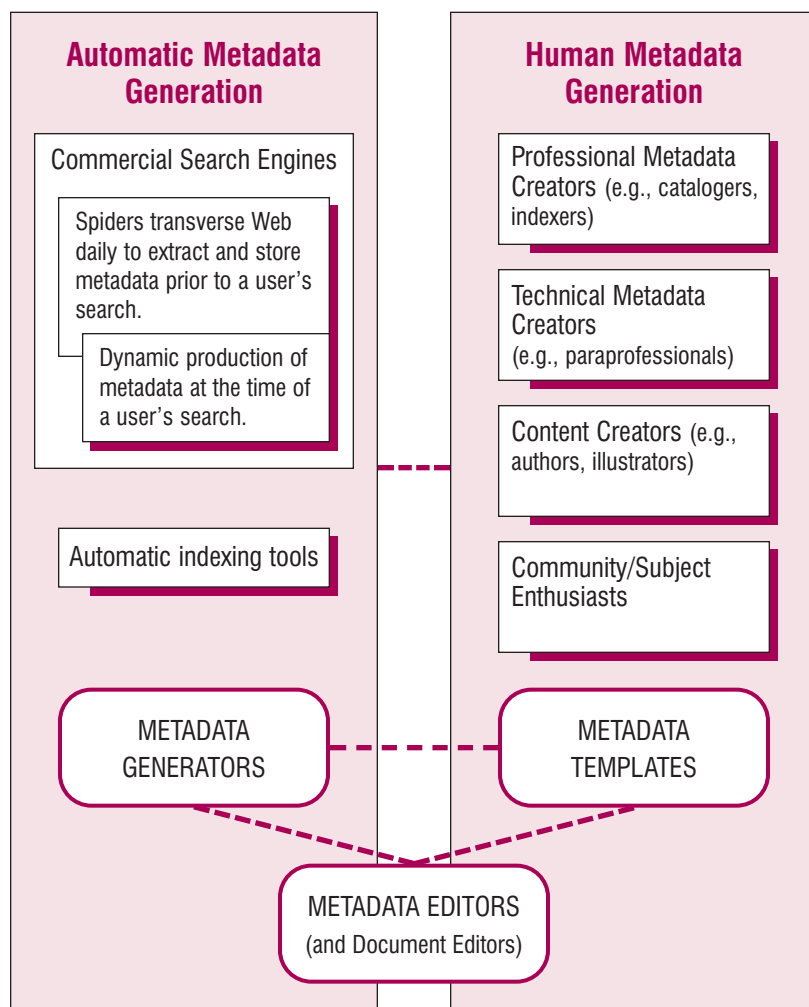
**Acknowledgement:** Portions of this article originally appeared in the *Encyclopedia of Library and Information Science* (Greenberg, J. (2002).

Metadata Platform) to support the creation and storage Dublin Core metadata within the Resource Description Framework (RDF) ([www.w3.org/TR/rdf-schema/](http://www.w3.org/TR/rdf-schema/)). The prevalence of manually generated metadata can, in part, be attributed to the fact that human metadata production is often superior to automatically produced metadata.

**Automatic metadata generation** depends on machine processing. Most familiar to the library and information science community is automatic indexing, which primarily focuses on a resource's subject content. Commercial search engines practice automatic metadata generation in two situations. First, metadata is produced automatically *prior to a user's search* by spiders that traverse the Web daily and extract and store resource metadata in the search engine's host database. A user's query is run first against this metadata store. Second, metadata is produced automatically and dynamically *at the time of the user's search* by executing a search and retrieval algorithm against the Web's global store of resources (beyond the search engine's host database). The second situation occurs when a user's query fails to match the metadata stored in the search engine's host database. Document representations in both situations are generally composed of the first few lines of a document (Web resource), locator information such as a uniform resource locator (URL) and title tag metadata.

If we liberally define objects for which metadata can be generated as "any entity, form or mode for which contextual data can be recorded" (see Greenberg, 2002, in **For Further Reading**), we find automatic generation operations taking place daily. Examples include the automatic generation of meta-

## Metadata Generation Framework



data on statements documenting an online purchase of an airline ticket, an automatic bank teller machine (ATM) transaction or telephone calls made during a previous month. Automatic processes permit human resources to be directed to more intellectually challenging metadata generation activities.

### Classes of Persons

Among the classes of persons involved in metadata generation, are *professional metadata creators*, *technical metadata creators*, *content creators* and *community or subject enthusiasts*. Although these classes of persons are defined separately here, for reasons of clarity, the distinctions are not always absolute.

**Professional metadata creators** include catalogers, indexers, Web masters and other persons who have had *high-level* train-

ing through a formal educational curriculum or an official on-the-job training program. Professional metadata creators have the intellectual capacity to make *sophisticated interpretative metadata-related decisions* and work with classificatory systems, complex schemas and other complex information standards. They are *third-party metadata creators* because they produce metadata for content created by other individuals. Given expert knowledge and skill, metadata professionals' greatest contributions may be in working with the more complex metadata schemas, instituting and overseeing a metadata production operation, instructing less skilled persons in metadata creation or helping to develop tools that facilitate metadata production.

**Technical metadata** creators include paraprofessionals, data in-putters and other persons who have had metadata training – but much less extensive training than the metadata professional. Technical creators are not expected to exercise discretion to the same degree as the metadata professionals, although they may take on more sophisticated tasks over time. They generally work with simpler metadata schemas and perform routine processes that are part of more complex metadata generation activities. For example, when a library orders a book, a library technical assistant, generally in the acquisition department, creates an “acquisition level” (acq level) MARC bibliographic record, which is a basic bibliographic description without authorized subject or name headings. When the book arrives in the library, a metadata professional uses the “acq level” metadata to create a full-level AACR2 MARC record – resulting in a richer and more standardized bibliographic description. It should be pointed out that the title of *technical* or *bibliographic* assistant can be misleading because frequently persons who are identified and paid as though they were technical metadata creators perform professional-like activities.

**Content creators** are individuals responsible for the creation of the intellectual content of a work. Researchers regularly produce abstracts, keywords and other types of metadata for their scientific and scholarly publications. Visual artists and crafts-persons sign and date their works. In the Web environment, content creators can provide metadata via a template or editor (see the *Tools* discussion below), while a webmaster “webifies” (makes the work Web-accessible). The National Digital Library of Theses and Dissertations (NDLTD) ([www.ndltd.org](http://www.ndltd.org)) and the Synthesis Coalition's National Engineering Education Delivery System (NEEDS) digital library for engi-

neering education ([www.needs.org/needs/index.jhtml](http://www.needs.org/needs/index.jhtml)) are digital library projects supporting author-generated metadata. Exploratory research has in fact shown that authors can produce fairly good metadata [Cruz & Krichel (2000); Greenberg et al. (2001)]. Facilitating content creator generated metadata makes sense when weighing the rapid growth of the Internet against the economics of hiring metadata professionals.

**Community or subject enthusiasts** have not had any formal metadata creation training, but have special subject knowledge and want to assist with documentation. A rudimentary view of this activity is found on personal Web pages that roughly classify, hyperlink to and/or provide other limited metadata for Web pages that document a topic of passion or interest. See topics such as “golden retrievers” (<http://home.att.net/~johnwaf/>) or “Italy and things Italian” ([www.geocities.com/Alejna2004/Alejnas\\_italy.html](http://www.geocities.com/Alejna2004/Alejnas_italy.html)).

The Fine Arts Museums of San Francisco’s *Thinker ImageBase* ([www.thinker.org/fam](http://www.thinker.org/fam)) provides an interesting and more formal example involving enthusiasts. Initiated during the Legion of Honor renovation following the Loma Prieta earthquake, *ImageBase* contains images and corresponding metadata for objects from the collections of the Fine Arts Museums of San Francisco (the de Young Museum and the Legion of Honor). Through a collaborative arrangement, *museum staff* provided artist’s name, date of creation, technique and other types of official museum registration metadata, and *community enthusiasts* (volunteers) assigned keywords to approximately 20,000 images. Community enthusiasts were used for two key reasons – to assist museum staff with object documentation and to enhance access through the provision of additional subject terms. Collaborative efforts between metadata professionals and community enthusiasts, as demonstrated by the *ImageBase* project, while not the norm, may become more common over time. The Web’s connectivity may increase collaborative metadata generation among other classes of persons as well.

## Tools

Metadata generation is supported by the following types of tools:

- **Human beings:** intellectual tools with the capacity to exercise discretion and perform data input
- **Standards & documentation:** metadata specifications, content guidelines, thesauri, classification lists and other types of standards and documentation that guide metadata creation
- **Devices:** technical compilations that “capture” and “store” metadata for either a database or resource header (e.g., the header of an HTML or XML document). Devices, as defined below, include *templates*, *editors* and *generators*.

**Templates** are basic *cribsheets* that sketch a framework or provide an outline of schema elements without linking to supporting documentation. Templates, in both print and electronic format, have been predominant in metadata generation, prob-

ably because they are simple to produce and maintain. These tools guide metadata creation through the provision of a *form* without the bells and whistles. An example is the Linux Software Map (LSM) Entry Template (<ftp://ftp.execpc.com/pub/lsm/LSM.README>) for metadata about Linux software packages. Guidelines associated with the LSM schema refer to the RFC822 standard for author name content syntax, among other standards, but the official template provides no linking mechanisms. Persons using this template generally work in a text editor, seek standards documentation on their own and submit their LSM records to a Linux repository via the File Transfer Protocol (FTP). The MARC bibliographic form supporting cataloging in many second-generation online catalogs has functioned in much the same way, without any sort of automatic linking to authority files and content guidelines. This facility is fortunately changing as many catalogs become Web-based and hyperlink to cataloging documentation, thus functioning more like an *editor*.

**Editors** are similar to templates in that they require human input. They are more sophisticated in that they provide direct access to standards and documentation underlying metadata

## Selected Metadata Generation Research Projects

- **Breaking the Metadata Generation Bottleneck**, School of Information Studies, National Science Foundation ([www.cnlp.org/research/project.asp?recid=6](http://www.cnlp.org/research/project.asp?recid=6)), is utilizing natural language processing and machine learning technologies to automate the assignment of metatags to educational resources in math and science.
- **Computational Linguistics for Metadata Building (CLiMB) Project**, Center for Research on Information Access, Columbia University Libraries ([www.columbia.edu/cu/cria/climb/](http://www.columbia.edu/cu/cria/climb/)), is using the latest developments in natural language processing to study problems of automatically extracting metadata from text.
- **GEM (Gateway to Educational Materials) Research** ([www.geminfo.org/Research/](http://www.geminfo.org/Research/)) is exploring the use of automatic indexing based on GEM metadata for major search elements to provide a mechanism for machine derivation of GEM metadata as a fairly precise “rough cut” metadata for resource discovery and retrieval.
- **Metadata Generation Research Project**, School of Information and Library Science, University of North Carolina (SILS/UNC-CH), in collaboration with the National Institute of Environmental Sciences (NIEHS), an Institute of the National Institutes of Health (NIH) (<http://ils.unc.edu/~janeg/mgr>), is developing a model to facilitate the most efficient and effective means of metadata production by integrating human and automatic processes in scientific research centers.
- **Tools for Information Resource Discovery on the World Wide Web**, Charlotte Jenkins, research associated with the further development of the Wolverhampton Web Library (WWLib) ([www.scit.wlv.ac.uk/~ex1253/research.html](http://www.scit.wlv.ac.uk/~ex1253/research.html)), is developing the “Classifier” – a tool that automatically classifies documents using the Dewey Decimal Classification within RDF.

creation. These tools often assist with syntactical aspects of metadata creation via automatic means. One of most popular Dublin Core editors is the Nordic Dublin Core Metadata Template ([www.lub.lu.se/cgi-bin/nmdc.pl](http://www.lub.lu.se/cgi-bin/nmdc.pl)). Although the term *template* appears in the official name of the Nordic Dublin Core Metadata Template, it is an editor according to the definitions offered in this article. This editor supports the generation of metadata records with HTML META tags for embedding in the header of a Web resource. The Nordic Template has been adapted to many different Dublin Core projects. A partial list of them is at: <http://dublincore.org/tools/>. Another example is the Reggie Metadata Editor (<http://metadata.net/dstc/>), which allows for metadata to be generated within RDF. Editors can also include off-the-shelf software like Metabrowser (<http://metabrowser.spirit.net.au/>), which hyperlinks to documentation supporting metadata generation and automatically provides the correct syntactical encoding. People work with a wide variety of Web forms when joining an organization, posting information on an online community bulletin board or purchasing a product over the Internet. All of these forms function as metadata editors documenting *transactions, activities, events* and other types of objects beyond the traditional information resource.

**Generators** support automatic metadata production. Note that the distinction given in this article between *editors* and *generators* is based loosely on those found under the “tools” link on the Meta Matters (Website) produced by the National Library of Australia ([www.nla.gov.au/meta/](http://www.nla.gov.au/meta/)). In the context of the Web, generators first require the submission of a uniform resource locator (URL), a persistent uniform resource identifier (PURL) or another Web address in order to *locate* the object. An algorithm is then used to *comb* an object’s content, including its source code, and automatically assign metadata. An example is found with the DC.dot generator ([www.ukoln.ac.uk/metadata/dcdot/](http://www.ukoln.ac.uk/metadata/dcdot/)), which requires the submission of a URL to locate and scan a resource’s content. This generator then automatically produces a Dublin Core record within RDF. The metadata can be embedded in the header of a XHTML or XML document or stored in a database. DC.dot supports metadata generation according to a number of different metadata schemas. Full “schema-specific” generators are fairly experimental because they can produce moderately accurate metadata for some elements such as the date a resource was last updated or MIME type, but results vary greatly for more intellectually demanding metadata such as subject descriptors.

One approach to dealing with the experimental and unpredictable nature of generators has been the creation of **hybrid metadata tools** that combine aspects of both editors and generators. DC.dot functions this way in that an editor permits a person to edit the metadata that is automatically generated. **Document editors** like Microsoft’s Front Page and Dreamweave exhibit the features of a hybrid tool by supporting human metadata generation of certain elements such as “key-

## For Further Reading

- Cruz, J. & Krichel, T. (2000). Cataloging economics preprints: An introduction to the RePEc Project. *Journal of Internet Cataloging*, 227-242.
- Greenberg, J. (2002). Metadata and the World Wide Web. In *The encyclopedia of library and information science* (Vol. 72, pp. 244-261). New York: Marcel Dekker.
- Greenberg, J., Pattuelli, M. C., Parsia, B., & W. D. Robertson. (2001). Author-generated Dublin Core Metadata for Web resources: A baseline study in an organization. *Journal of Digital Information (JoDI)*: <http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Greenberg/>

words” and “description” and automatically producing other metadata as part of the document-creation process such as “date document was produced” and “document format.”

## A Research Context

This overview has identified *processes, persons* and *tools* comprising a metadata generation framework. Although we have commented on the strengths and weaknesses of these means and mechanisms, the discussion has served more to sketch the framework than to present empirical results. Decisions about the processes, persons and tools to employ for metadata generation depend on a project’s architecture, complexity of desired metadata schema, time allotment and project deliverables and the availability of human, financial and time resources. Clearly, different combinations of these metadata generation components will be more effective in different environments. Research efforts testing various combinations of processes, people and tools will help establish useful models to guide metadata generation activities. A list of selected research projects exploring aspects of metadata generation is found in the sidebar.

This article concludes by highlighting the Metadata Generation Research Project (<http://ils.unc.edu/~janeg/mgr>) being conducted at the School of Information and Library Science, University of North Carolina (SILS/UNC-CH), in collaboration with the National Institute of Environmental Sciences (NIEHS), an Institute of the National Institutes of Health (NIH). The Metadata Generation Research Project was launched with funding from Microsoft Research and is continuing with support from the OCLC, Online Computing Library Center. Research goals underlying this project include studying human and automatic metadata generation processes, developing protocols for collaboration between resource authors (content creators) and metadata professionals during metadata generation, evaluating the integration of collaborative human metadata generation processes with automatic generation processes and considering implications for the development of the Semantic Web. A model is being developed to facilitate the most efficient and effective means of metadata production in scientific research centers. The model and the underlying methods of inquiry will, in turn, aid future metadata generation investigations. In closing, this work, along with the other initiatives listed, will ultimately aid in efficient generation of good quality metadata and in making the Web’s vast collection of rich resources more accessible.

# Data Collection for Controlled Vocabulary Interoperability—Dublin Core Audience Element

by Joseph T. Tennis

---

*Joseph Tennis is a doctoral student in the Information School of the University of Washington. He can be reached by e-mail at [jtennis@u.washington.edu](mailto:jtennis@u.washington.edu)*

---

This paper outlines the assumptions, process and results of a pilot study of issues of interoperability among a set of seven existing controlled vocabulary schemes that make statements about the audience of an educational resource. The notion of *audience* for the study was defined in terms of the semantics of the Dublin Core metadata element of the same name: “A category of user for whom the resource is intended.” The study used a data collection technique, card sorting, to see how non-expert users (1) sorted terms in the seven vocabularies into relationships and (2) what their thought processes were in sorting these terms.

The need for controlled vocabulary interoperability is a pressing concern for the education community as well as many others. In particular, the current study was informed by the need of the Dublin Core Education Working Group ([www.dublincore.org/groups/education/](http://www.dublincore.org/groups/education/)) to explore the possibility of a high-level switching language in an application profile for the Dublin Core Metadata Initiative (DCMI) audience element. An abundance of educational resources exists, many of which are available in the networked environment. Yet, there are various conceptualizations of the domain in the form of different controlled vocabularies that limit access. Controlled vocabulary interoperability would allow these different conceptualizations to remain intact, thereby serving local needs while allowing users to navigate across collections and exploiting the intellectual network of resources available.

## Assumptions of Controlled Vocabulary Interoperability

Controlled vocabulary interoperability is a problem that is complex and has a long history in information science. From the 1960s onward there are discussions in the literature of attempts to reconcile at least two controlled vocabularies from the same domain. The result of this body of research has been the creation of mapping and switching techniques. Mapping attempts to match one-for-one the terms of each controlled vocabulary. This technique assumes that you can find the same meaning (in its extension and its intension) between two terms in two different controlled vocabularies. The other result, switching, advocates a third intervening language be constructed that can switch between the two controlled vocabularies. This also assumes that the problems of meaning can be reconciled between three terms: two controlled vocabulary terms and one switching term. Examples of these types of interoperability are given in Tennis (in press. See **For Further Reading**.)

Lancaster presents the classic argument against the perfect actualization of interoperability. Lancaster states that the structure of subject access systems confounds any seamless switching. The problems inherent in switching between two vocabularies are differences in (1) overlap of subject matter, (2) specificity, (3) degree of pre-coordination and (4) hierarchical, synonymous and other relationship structure. For Lancaster each term is

defined in relationship to other terms in the controlled vocabulary. As a consequence the meaning of the term may be more broadly or narrowly defined (1 and 2 above) or the placement of that term below another term may shape its meaning implicitly (3 and 4 above).

Therefore the system that allows interoperability between two (or more) controlled vocabularies must reduce or eliminate structure while maintaining meaning. By reducing or eliminating the structure of terms from controlled vocabularies, it may be possible to shift interoperability toward a pragmatic management of these terms as they exist for the user and the user's needs.

The questions and assumptions in interoperability are philosophical. What is the best approach? What is the best-reasoned design of an interoperable space? Interoperability can become solely an object of investigation and not necessarily an area of development of best practices and principles. Since the end goal of this study is to develop a technique or, perhaps, a set of techniques to support interoperability among existing controlled vocabularies used with education materials, the issues in the study were cast as a problem of information management that had to be informed through empirical analysis. Given this problem, what techniques can be employed to inform a decision about how to manage the interoperability between these education-focused controlled vocabularies? Specifically, what kind of data do we need to begin the iterative design of a term management space that will allow for interoperability? One data collection technique available is card sorting.

### Assumptions of Card Sorting

Card sorting, or just sorting, is used by various disciplines to examine how individuals organize a given set of cards. Often the cards have terms on them. Sometimes other materials are sorted. For example, Carlyle (2001) sorted different instances of one work. The results of the sort are interpreted as a user's (sorter's) conceptual arrangement of the given universe of objects. In information science and in usability work, sorting is a tool that can be used to help inform the design of online displays. Carlyle reviews various aspects of sorting. She also points to Fidel's suggestion that data derived from users should be used to inform systems design. Carlyle also suggests that various user groups, not just specialists in a field, should be part of the card sort.

It is with these assumptions that this pilot study began. We wanted to investigate the viability of a card sort technique to inform an aspect of systems design – specifically the design of an interoperability space between various controlled vocabularies used in education. The results would allow me to evaluate the method as a data collection technique for interoperability and to modify the card sort technique if necessary.

### The Card Sort

The card-sort pilot study was a first attempt at reconciling seven different controlled vocabularies in the field of edu-

cation. The terms sorted were audience types. Audience types are those people or groups that might use the resources in an education information store. For example, these people may use lesson plans, simulation tools or instructional texts. Audiences include parents, teachers, vocational educators, counselors, doctors, etc. Many are very general terms in themselves. Yet their meanings are restricted because they are deployed in the context of the education domain. Thus counselors are specific types of counselors – those whose work focuses on student activities and problems.

A purposive sample of terms was chosen from controlled vocabularies developed by Education Network Australia, European Treasury Browser, Gateway to Educational Materials, Instructional Management Systems, Australian Government Locator Service, UK National Curriculum and U.S. Department of Education.

The terms in the purposive sample were considered unique and more general than those left out of the sample. Due to the experimental nature of the card sort a limited set of 37 terms was selected (see box). The participants were asked to sort these cards into relationships that made sense to them. They were told that the terms were from the domain of education and represented kinds of users (audiences) of educational materials. The participants were also asked to label any groups or subgroups that resulted from their sorting. The purpose of this labeling was to explore potential high-level vocabulary terms that might be used to represent the groups or subgroups. While the participants were sorting, they were asked to talk aloud so that their thought processes could be better understood. For this pilot nine participants volunteered their time.

### Sort Terms

Adult Educators	Non-teaching Staff
Alumni	Political Decision Makers
Amateur	School Aides
Author	School Doctor
College/University Instructors	School Leadership
Counselors	School Nurses
Curriculum Supervisors	School Personnel Workers
Dropouts	School Psychologists
Educational Administration	School Publisher
Educationalists	Speech Therapist
Families	Stopouts
Graduates	Students
Guidance Officer	Teacher Interns
Inspector	Teachers
Learner	Technology Coordinators
Librarians	Trainer
Manager	Tutors
Managerial Staff	Vocational Educators
Media Specialists	

## Results of the Card Sort

This pilot card sort generated three types of data: (1) the piles of cards (with high-level vocabulary labels supplied by the participants); (2) a transcript of the talk-aloud; and (3) observations of the act of sorting that may prove quite instructive.

*Sort Results:* The piles of cards with their labels were entered into Microsoft Excel as occurrence matrices where the occurrence of a term with another term (to form a group) was indicated in the matrix with a "1." A term that did not appear with another term was marked with a "0." The matrices were totaled. A number of cluster analyses were conducted on the total. The cluster analysis describes the aggregate result of the sorting task by all nine participants. However, individually each pile informs us of details that are lost in the aggregate.

The labels provided by the participants should be examined conceptually rather than literally. The participants were encouraged to be casual through the talk-aloud protocol. This resulted in some labels for groups that reflected participant moods or humor. However, these labels still carry weight as grouping terms on the conceptual level. Two participant labels appear below. Keep in mind that the participants sorted all 37 cards into these groups and labeled them.

*Participant "A" Card Group Labels:* Education Consumers; Teachers; Administrators; Outside

*Participant "B" Card Group Labels:* School Administration; School Publishing; Teaching Staff; Students, Alumni and Families; Non-Teaching Staff

As can be seen in this data the two participants do not agree on the number of groups or the names of those groups. That is why it is helpful to see the results in the aggregate.

The aggregate picture of the card sort, when examined using a cluster analysis (Ward's Method in the SPSS program) looks like this:

*Cluster A:* Animateur; Media Specialists; School Aides; Librarians

*Cluster B:* Educational Administration; Political Decision Makers; School Publisher; Author

*Cluster C:* Inspector; School Leadership; Managerial Staff; Manager; School Personnel Worker; Curriculum Supervisor; Non-teaching Staff; Technology Coordinator

*Cluster D:* Guidance Officer; Speech Therapist; School Nurses; Counselors; School Psychologists; School Doctor

*Cluster E:* Graduates; Learner; Alumni; Stopouts; Dropouts; Students; Families

*Cluster F:* Teachers; Trainer; College/University Instructors; Educationalists; Tutors; Teacher Interns; Vocational Educators; Adult Educators

This aggregate fits much of the criteria for a good clustering of the terms. These terms fit first and foremost with a common-sense analysis. These terms do go together and they are similar to the individual participant sorts. However, the agglomeration schedule generated by the Wards Linkage

process of SPSS shows that the optimum number of clusters is four. There are six clusters in this data. The problem lies with groups A through C. The participants were very creative when sorting these terms – distributing them across many groups. They were not as creative when sorting D, E and F. As a consequence these terms appear most often together and, therefore, fall naturally together in the cluster analysis.

There is additional work that can be done with this data. Cluster analysis is full of variations on method and of controversy over applicability. In this study, cluster analysis was used as a descriptive technique to find structure in the data gathered. It was not in any way intended as a predictive analysis. Because of wide ranges of interpretations of cluster analysis, it is encouraging to see such strong evidence in the common-sense analysis (the coherence of the six groupings given above). It seems that this data could inform the design of an interoperability space.

## Results of the Talk-Aloud

The transcript of the talk-aloud is still being analyzed. However, from preliminary examinations of the data, it seems clear that the participants wanted to construct a meaning for a term *in relationship* to the other terms available. Often the participant was confused, expressed this confusion and made decisions without certainty.

## Results of the Observation

Across all participants common phenomena were observed. First, each of the participants laid the entire set of cards out before him or her. From this undivided universe of 37 cards, they began the process of sorting. A second common process in this sorting task was a vacillation between sorting from the bottom up and sorting from the top down. Some categories grew from lumping like things. Others were made from dividing the unlike groups of terms, then further dividing. However, neither approach (top-down nor bottom-up) was used exclusively by any of the participants.

## Conclusion – Data collection and Interoperability

As a data collection technique, card sorting provides the researcher with various types of data – some of it very rich. As part of the decision making process for systems design, a card sort should be only one of many sources of information. Collecting data from card-sort techniques will not solve the philosophical problems of controlled vocabulary interoperability. As Miller notes (2000) much work needs to be done to understand what is wanted from interoperability in general. From a menu design perspective, the pragmatic management perspective, card sort techniques allow those in need of solutions a way to inform their decisions, as McDonald and Schaneveldt show (1988).

More research is needed into the particularities of card sort data collection in this venue. What role does the expertise of the sorter play in the resulting piles? Hayhoe (1990)

## Cover Stories

suggests that it is beneficial to have both novices and experts shape the resulting aggregate for cluster analysis. How do terms within the domain of education like *Animateur*, unfamiliar to some, influence the analysis? In this particular case, the talk-aloud and observation data helped to identify the problems the participants had with *Animateur*, whereas the aggregate analysis of all of the piles hid this problem.

From a pragmatic point of view, it is necessary to test this data collection technique a number of times and in different contexts. If card sorting is to become a viable method for aiding controlled vocabulary interoperability it should be rigorous yet easy to perform. It should generate enough data to give a coherent picture, but not create paralysis from data overload.

Following a close analysis of this pilot study, we hope to design and carry out a more robust research project. There is much work to be done in controlled vocabulary interoperability if metadata initiatives are to take advantage of the networked environment. Linking intellectually as well as *hypertextually* to stores of information will require interoperability. In order to ensure interoperability those who work with metadata will want best practices information about how to construct an interoperability space from different controlled vocabularies.

## For Further Reading

### Interoperability

Miller, P. (June 2000). Interoperability. "What is it and why should I want it?" *Ariadne*, 24. Available: [www.ariadne.ac.uk/issue24/interoperability/intro.html](http://www.ariadne.ac.uk/issue24/interoperability/intro.html).

### Vocabulary Mapping and Switching Languages

Dahlberg, I. (1996). Compatibility and integration of order systems 1960-1995: An annotated bibliography. In *Compatibility and Integration of Order Systems: Research Seminar Proceedings of the TIP/ISKO Meeting, Warsaw, 13-15 September, 1995*. Warsaw: Wydawnictwo.

Lancaster, F. W. (1986). *Vocabulary control for information retrieval*. (2nd ed.) Arlington, VA: Information Resources Press.

Tennis, J. T. (in press). Layers of meaning: Disentangling subject access interoperability. In Efthimiadis, E. (Ed.) *Advances in Classification Research*, 12.

### Sorting and User-Centered Design

Carlyle, A. (2001). Developing organized information displays for voluminous works: a study of user clustering behavior. *Information Processing and Management*, 37, 677-699.

Fidel, R. (1994). User-centered indexing. *Journal of the American Society for Information Science*, 45, 572-576.

Hayhoe, D. (1990). Sorting-based menu categories. *International Journal of Man-Machine Studies*, 33, 677-705.

McDonald, J. E. & Schanefeldt, R. W. (1988). The application of user knowledge to interface design. In Guindon, R. (Ed) *Cognitive science and its application to human-computer interaction* (pp. 89-338). Hillsdale, NJ: Lawrence Erlbaum, pp. 89-338.

The **BULLETIN OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY** is a **BIMONTHLY PUBLICATION** that serves as the newsletter of the Society. It publishes short articles on a **BROAD RANGE OF TOPICS** of current concern to **ASIST MEMBERS**, focusing particularly on material of interest to practitioners. Readers are **ENCOURAGED TO SUGGEST** topics of interest or alert the Editor of suitable material that may have been presented at ASIST-sponsored events or elsewhere. In addition, authors are **ENCOURAGED TO SUBMIT** articles on topics such as **CURRENT PRACTICE, PUBLIC POLICY, LEGISLATION, STANDARDS, PILOT PROJECTS, STATE-OF-THE ART REVIEWS** or **OVERVIEWS OF EVOLVING TECHNOLOGY AND ITS IMPACT**. Articles informing the membership about various developments within ASIST are very welcome, as are articles reporting on **ACTIVITIES OUTSIDE THE UNITED STATES**. The *Bulletin* encourages original articles, but will consider **TIMELY MATERIAL** that has been presented or published elsewhere. Articles are posted in full on the **ASIS Web Site** at <http://www.asis.org/Bulletin/index.html>

Authors interested in developing material for a focused issue are urged to contact the Editor directly.

Authors are encouraged to discuss article ideas with the Editor if there are questions about suitability or relevance.

**Irene L. Travis, Editor**  
*Bulletin of the American Society for Information Science and Technology*  
1320 Fenwick Lane,  
Silver Spring, MD 20910  
(301) 495-0900  
[Bulletin@asis.org](mailto:Bulletin@asis.org)

of the American Society for Information Science and Technology  
**BULLETIN**

# A Knowledge Network Constructed by Integrating Classification, Thesaurus and Metadata in a Digital Library

by Wang Jun

Wang Jun is associated with the Information Management Department of Peking University, Beijing, China, and can be reached by e-mail at [junwang@pku.edu.cn](mailto:junwang@pku.edu.cn)

**Editor's note:** This article has been condensed from the paper that was awarded third place in the ASIST SIG/III 2002 International Paper Competition. Other papers from this year's contest, including the first and second prize winners, will appear in later issues of the *Bulletin*.

**K**nowledge management in digital libraries is a universal problem. Keyword-based searching is applied everywhere no matter whether the resources are indexed databases or full-text Web pages. In keyword matching, the valuable content description and indexing of the metadata, such as the subject descriptors and the classification notations, are merely treated as common keywords to be matched with the user query. Without the support of vocabulary control tools, such as classification systems and thesauri, the intelligent labor of content analysis, description and indexing in metadata production are seriously wasted. New retrieval paradigms are needed to exploit the potential of the metadata resources. Could classification and thesauri, which contain the condensed intelligence of generations of librarians, be used in a digital library to organize the networked information, especially metadata, to facilitate their usability and change the digital library into a knowledge management environment?

To examine that question, we designed and implemented a new paradigm that incorporates a classification system, a thesaurus and metadata. The classification and the thesaurus are merged into a concept network, and the metadata are distributed into the nodes of the concept network according to their subjects. The abstract concept node instantiated with the related metadata records becomes a knowledge node. A coherent and consistent knowledge network is thus formed. It is not only a framework for resource organization but also a structure for knowledge navigation, retrieval and learning.

We have built an experimental system based on the *Chinese Classification and Thesaurus*, which is the most comprehensive and authoritative in China, and we have incorporated more than 5000 bibliographic records in the computing domain from the Peking University Library. The result is encouraging. In this article, we review the tools, the architecture and the implementation of our experimental system, which is called Vision.

## The Development of the Chinese Classification and Thesaurus

China has a long tradition of using classification due to her abundant ancient books. Modern Chinese classification was greatly influenced by Dewey although the *Dewey Decimal Classification* wasn't popular in China. All the classifications now in use were created after the foundation of the People's Republic of China. The *Book Classification of Chinese Libraries (BCCL)*, which was published first in 1975 and has undergone four revisions, is the most developed.

The most famous comprehensive thesaurus in China is the *Chinese Thesaurus (CT)*, compiled between 1974 and 1980 in an effort involving more than 1000 people. It was the biggest thesaurus at that time, containing 91,158 preferred terms and 17,410 non-preferred terms. Influenced by faceted thesauri, a huge project was started in 1986 to combine the *BCCL* and the *CT* into the *Chinese Classification and Thesaurus (CCT)*. More than 40 institutions were involved, and it was finished in 1994 and contains 14 million words in six volumes. Currently, it is used in all public libraries and more

than 90 percent of non-public libraries and information institutions of China.

The *CCT* is not designed for the network environment and has seldom been applied there. Its application has several inherent obstacles, most of which, as Zhang Qiyu recently explained [“Discussions of the information retrieval language of 21 Century,” *Forum of Libraries*, 21 (5)], are common to other classifications and thesauri of broad scope. These problems include currency and difficulty in tailoring coverage to specific domains. Moreover, the tools were designed for the organization of information resources and for the management of hard copy documents, not for information retrieval. They are very complex and require that indexers and classifiers have extensive training. Finally, in the case of the *CCT* the classification and thesaurus are relatively independent of each other and cannot be updated synchronously.

### The Knowledge Network

The exceptions, of course, are the online public access systems (OPACs) in Chinese libraries. Their bibliographic data are indexed strictly according to the *CCT*, and as collections of living materials they contain plenty of new professional terms in their title fields. To overcome the obstacles mentioned above, the classification, the thesaurus and the bibliographic data can be combined to complement each other. The knowledge structure of the classification and thesaurus provides a skeleton for the organization of the bibliographic data; the concrete bibliographic data restore blood and flesh to the skeleton. New terms can be extracted automatically from the bibliographic data to update the classification and thesaurus, which is based on the mapping between the subject descriptions and the titles that they index; while the classification and thesaurus are customized to the specific domains of the OPAC resources. A knowledge network thus formed provides the user with a natural structure for navigation, searching and learning. We will call it a KNICTM (Knowledge Network Integrated of Classification, Thesaurus and Metadata).

In *Vision* we combined the classification and indexing terms from the computing domain in the *CCT* with all the bibliographic records for Chinese materials in computer science held by the Peking University Library published between 1990 and 1999, which provided a database of more than 5000 bibliographic records for our *Vision* system.

A KNICTM is built in three steps:

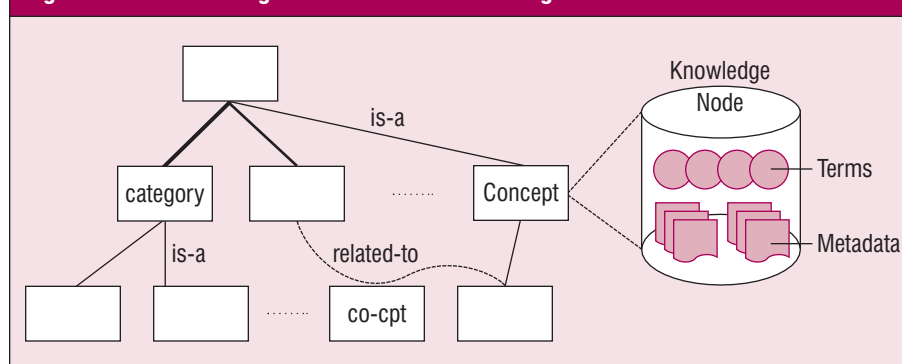
1. *Construction of the original concept nodes based on the classification and thesaurus.* First, the thesaurus is turned into a concept network consisting of nodes and edges. A node

is composed of the synonym set of a subject descriptor, including the descriptor and all the terms connected to it by the equivalence relationship (Use/Use For) in the thesaurus. If there is a hierarchical relationship (Broader Term/ Narrower Term) between two terms, an “is-a” edge is set up between the corresponding concept nodes. Next, the classification scheme is embedded in this original concept network as a discipline-oriented hierarchical backbone (Figure 1). Since the *CCT* is a reciprocal index between the *BCCL* and the *CT*, rather than a faceted thesaurus, there is no direct mapping between the categories of the *BCCL* and the concepts of the *CT*. Therefore, category nodes had to be created, and the relationships established among the category nodes and the concept nodes.

2. *Distribution of the bibliographic data to the concept network.* The bibliographic data are arranged into the nodes of the original concept network according to their subjects. This is the key task of the KNICTM construction. Supported by the bibliographic data, each abstract concept node becomes a knowledge node where the abstract concept is bound to the metadata records. And the concept network turns into a knowledge network formed by the integration of classification, thesaurus and metadata. It is a kind of metadata “shelving.” If a bibliographic record contains only one subject descriptor, we take the record as one of the instances of the corresponding concept node and add the record to the node. If it contains several descriptors, then we add it into all the related concept nodes as instances of them. If it contains a composite subject described by a coordination of descriptors, we create a new concept node, and connect the node to all the corresponding nodes of the coordinate descriptors with “related-to” edges. The newly created concept node is called a co-concept node and has bibliographic records only and no term for the moment.

For example, a bibliographic record with the title

Figure 1. The knowledge network with a knowledge node zoomed in.



“Internet Firewall Technologies” is indexed with the string “Network--Security” provided that there is no “Firewall” in the thesaurus. To add this record into the concept network, a co-concept node is created and connected to the concept nodes for “Network” and “Security” by “related-to” edges. Since the associative relationships easily get out of control in a thesaurus, we don’t create them in the Step 1. Only when the correlation of two concepts is supported by a bibliographic record do we establish the “related-to” relationship between them through a co-concept. Thus the bibliographic data function as the verifications of the associative relationship. In Step 3 when a new extracted term with the meaning of the co-concept occurs – “Firewall” in the above example – the new term is added to the co-concept. The KNICTM needs manual examination periodically to confirm the co-concepts created. When a preferred term is determined for the co-concept, the co-concept node becomes a common concept node.

3. *Enhancement of the KNICTM.* The last and the most difficult task is to mine new terms from the metadata collection to enhance the KNICTM. The title of a scientific document usually summarizes its content and reveals its central topics. A direct mapping exists between the keywords of the title and the subject descriptors and the classification notation used to index the document. Based on this mapping, statistic and semantic techniques can be applied to extract new terms from the title and add them into the concept network. There are three difficulties:
  - Segmenting the title into words and phrases – a classical problem for Chinese.
  - Extracting valuable terms from the common terms in title.
  - Determining the point where the extracted terms should be inserted into the KNICTM.

### The Benefits of the KNICTM

The KNICTM provides a number of benefits:

- *A framework for the organization of network resources.* It is a network of knowledge with substantial data appended rather than a mere abstract concept network. As the instances of the concept, the metadata records inherit all the relationships among the concepts. The metadata records which were isolated from each other become semantically connected now and are woven into an interconnected knowledge network.
- *An adaptive concept network based on the applied resources.* The classification and thesaurus are the representation of general knowledge and cannot fit a specific information collection perfectly. The KNICTM is an adaptive concept network capable of self-cus-

tomizing based on the scale and domain of the given collection. The nodes and edges, supported by the metadata instances, prove the usability of the corresponding concepts. If nodes and edges have no metadata instances the corresponding concepts and relationships are unusable and may need updating. Furthermore, statistic and semantic techniques can be applied in mining new terms, concepts and relationships in the metadata collection to enrich the concept network automatically.

- *A structure for knowledge navigation and retrieval.* Keyword-based search seriously under-exploits the value of the metadata. The KNICTM provides a conceptual retrieval network and visual navigational ontology. First, the KNICTM can guide a user to clarify the information demand and express a query clearly. Second, because all the metadata have been arranged into the KNICTM, there is no need for the user to dig into the metadata collection by keyword matching. It is only necessary to locate the knowledge node that best matches the query and follow the surrounding edges to reach other nodes to complete the process. Third, now that all the metadata have been arranged into the structure of the KNICTM according to their subjects, the retrieval result is displayed in that structure, already ranked and classified.
- *A well-organized knowledge network to support knowledge learning.* The knowledge nodes are organized into a discipline-based hierarchy and clustered into topic areas through the links among them. A friendly interface like the Cat-a-Cone developed by Marti Hearst and Chandu Karadi can display the organization of the knowledge nodes. A user facilitated by such an interface can learn the discipline structure of a domain, master the professional terms, understand the relationships among the subjects and pick up the documents to study.
- *A digital library of knowledge management.* The most essential elements of a library are its information resources and the classification and thesauri, which are its information organization and retrieval tools. In KNICTM, these elements have been integrated into a coherent and consistent knowledge network. And the KNICTM could be easily extended to support other activities of digital libraries, such as collecting and indexing. Thus all the activities of the digital library, including indexing, organization, navigation, retrieval and learning, could center this knowledge network. If it can develop continuously, the KNICTM will bring the digital library from information management to knowledge management.

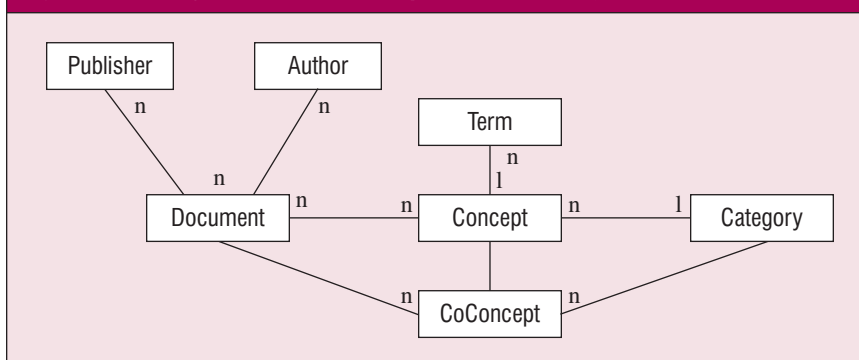
### The Construction of the Knowledge Network

We have completed the first phase of the Vision system. It has a client/server architecture. On the server side the knowl-

edge network is supported by Oracle9i. On the client side is a user interface implemented in Java. We chose Oracle9i for its powerful object-oriented features, such as nested tables and variable arrays, which support our complex objects. Java makes it easy to transfer the system to the Web.

*The Ontology Design.* There are many objects in our system and their relationships are complex. We therefore used ontology tools such as *Ontolingua* and *Protégé* to design it. We then converted this ontology into the database schema. Our ontology consists of seven classes: term, concept, co-concept, category, document, author and publisher. Their names reflect their meanings, but the relationships among them are entangled. Figure 2 depicts these relationships. The numbers indicate the cardinality of the links. We converted the completed ontology into the relational schema of the database system and created the corresponding tables in Oracle.

Figure 2. The objects and relationships in the Vision system.



*The Server Side: The Knowledge Network.* The original dataset used to build the Vision system included the e-text of the CCT and the bibliographic data of the computing domain. Both of them were provided by the Peking University Library. The characteristics of the original data had considerable influence on the system design and implementation.

There are three steps in building the Vision server:

- *The e-text file of the CCT is processed to set up the fundamental structure of the Vision system.* A particular tool was developed to serve this purpose. The e-text of the CCT is read in and all the entries (categories and terms) on computer science are processed. According to the structure, layout and notation rules of these entries, the related records are created and appended into four tables respectively: TERM, CONCEPT, CoCONCEPT and CATEGORY. Through this process, we collected 2194 terms, including 1684 preferred terms which became concepts, and 278 non-preferred terms. Some non-computing-domain terms were also captured since they are the related terms.

- *The bibliographic data are loaded in the database and organized into the original concept network constructed in the preceding process.* The bibliographic data are in CNMARC format. We developed a tool to decode the CNMARC format and extract the required fields (title, subject, author, etc.) to form a new record, which is appended to the table DOCUMENT. In total 5053 document records were created. Others were discarded for various reasons, for example, unrecognized title or two ISBN numbers. Such data processing required a lot of time and energy. After the data was loaded, the records of the DOCUMENT table were connected with the records of the CONCEPT table based on the correspondence. When necessary, a new record was created in CoCONCEPT table. These processes accomplish the task of organizing the metadata into the knowledge network described above.

- *New terms are extracted from the DOCUMENT table and added to enhance the knowledge network of the Vision system.* Some of the problems have already been mentioned. This process is the focus of the ongoing second phase of the Vision project, so here we just outline roughly what we have done to date.

- *Extraction:* At present a statistical algorithm is applied to extract terms in titles. First, the title is segmented into basic words and phrases using a general segmentation tool, and then the co-occurrence frequencies of neighboring terms are counted. If the frequency is higher than a given threshold, the combination is selected as a candidate term. Then we look at the distribution of subject categories in the set of documents in which the candidate term occurs. If the distribution of the subject categories is convergent, the new term is accepted.
- *Insertion:* The convergent point found above helps to determine the position where the new term should be inserted. We are considering applying Lattice Theory or Formal Concept Analysis to this problem.

*The Client Side: Knowledge Navigation and Retrieval.* We implemented a system in Java to navigate and retrieve the Vision knowledge network. Figure 3 is a snapshot of the user interface. There are four physical areas in the interface: the query dialog, the concept network window, the information window and the document window.

Within the concept window, there are three basic ways to view the concept network: hierarchical tree, alphabetical list and concept family. The hierarchical tree is similar to a faceted thesaurus. All the categories and concepts (identified by the

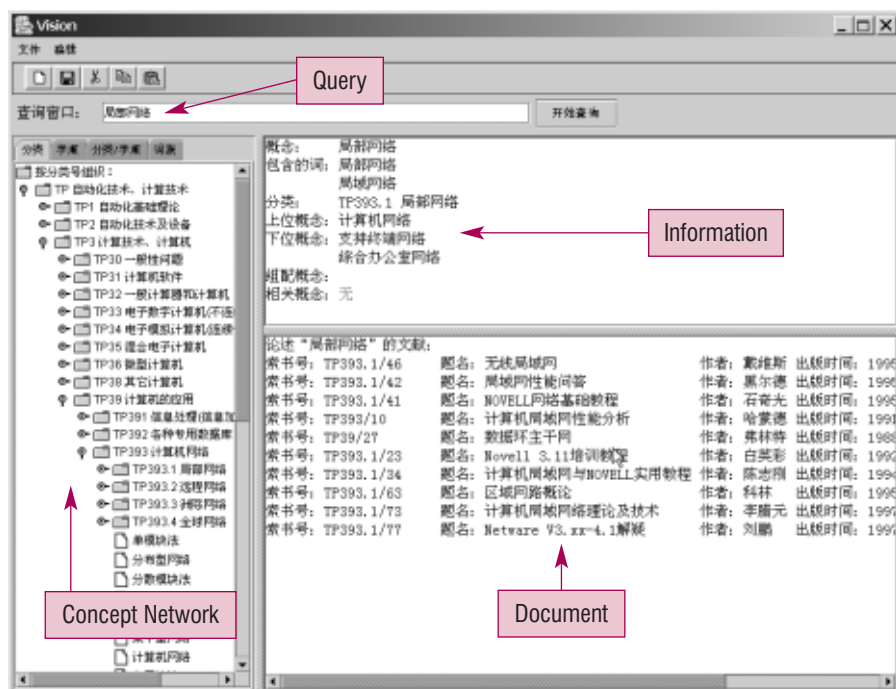


Figure 3. The interface of the Vision system

preferred terms) are organized into an expandable conceptual tree. The alphabetic list is an index of all the terms in alphabetical (Chinese Pin Yin) order. The concepts can also be organized into concept families, that is, the term families of the thesaurus, and listed in alphabetical order by the top concepts. There is a fourth option, which is a hybrid of the hierarchical tree and the alphabetic list.

When the user clicks on a concept, its detailed information is displayed in the information window, including its term set, super-concept, sub-concepts, corresponding category and the co-concepts around it. The documents connected with it are displayed in the document window. All the windows trigger each other and act in a chain, and all the objects in the windows are clickable.

## Conclusion and Future Work

Centuries of library work have proved that the organization of information is the basis for the sufficient utilization of information resources. It's the fundamental value of library. The same is true for digital libraries. For lack of organization the potential of metadata as one of the most important networked resources is not exploited sufficiently. This article has presented an approach to organizing metadata into an integrated knowledge network and setting up a new paradigm for knowledge management in digital libraries. Our approach is differentiated from other ontology-driven and concept-based systems by its incorporation of concepts and the relevant metadata records into integrated knowledge nodes that form a

knowledge network. Our experiment also demonstrates that the traditional resources such as bibliographic data still have indispensable values worth further exploration in spite of the continuous increase in various digital resources.

The Vision system is entering its second phase. We are endeavoring to achieve the following goals:

- A better method to compute the extension and intension of a term and determine its position in the concept network. This is critical if the concept network is to be a self-sufficient system. We are now considering applying a modified version of Formal Concept Analysis to this problem.
- A language for concept query and manipulation that will simplify operations on the concept network and add an automatic query expansion and contraction mechanism.
- A visualization interface such as

Cat-a-Cone or InxightStar-Tree, which can provide more friendly interaction with the user. A visualization interface that is structurally isomorphic to the concept network will support knowledge learning more powerfully.

When all the aspects of the system have been tested, the system will be translated to the Web and incorporated into the current OPAC system.

To integrate classification, a thesaurus and metadata into a coherent knowledge network has promising applications in digital libraries. It could easily be expanded to support automatic classification and indexing in scientific domains. Enhanced by the bibliographic data, the knowledge network could absorb other metadata, such as the index databases of journals, magazines or newspapers.

The Web community also recognizes the importance of the standardization and organization of the Web information. XML, RDF, Dublin Core and other specifications are preparing the Web for the manageable Web – the Semantic Web as envisioned by Berners-Lee ([www.w3.org/2000/Talks/1206xml2k-tbl](http://www.w3.org/2000/Talks/1206xml2k-tbl)). But how to construct it? Our paradigm provides one approach.

## Acknowledgements

This research is a portion of my doctoral dissertation at Peking University. I wish particularly to thank Deng Peng, Zhu Xingguo and Zu Yong who assisted me in developing the Vision system, and I'd like to thank Professors Yang Dongqin and Tang Shiwei as well as Dai Longji, president of the Peking University Library, and his staff.

# On Trust and Users

by Andrew Dillon

*Andrew Dillon is dean, GSLIS, University of Texas at Austin, and can be reached by e-mail at [adillon@gslis.utexas.edu](mailto:adillon@gslis.utexas.edu).*

I've found myself thinking a lot about trust recently: trust in computers, trust in users and trust in information architects. At first these were just vague rumblings, a concern that all was not quite right, provoked in part by the comments and responses I noted on various lists where otherwise bright people seemed to lose their balance over certain topics (such as "who can call themselves an IA?"). But it came to a head when I read a blog by Kevin Marks (<http://epeus.blogspot.com/>) entitled "Trust people, not computers." Apparently *The Economist* ran a report on a study from IBM that shows Internet adoption in a culture has less to do with the classic indices of number of telephone lines or years of education and more to do with the cultural norms related to trust in dealing with others. While the notion of trust has received most attention in work on e-commerce, or academic writing on authenticity and authority, such work is highly localized. Here was a broad sweeping idea about dispositions towards information technology that transcended narrow contexts.

As I write this column I am teaching a class on user cognition and behavior that explores high-level principles governing how users act in information environments. You might think this field has established more than a few such principles over the years. However the process of teaching, as so often happens, has caused me to question many of the assumptions I have made about the trust we can place in users and the information professionals who claim to design

for and serve them. It seems to me that the basic ideas of user-centeredness, that we gain reliable and valid design guidance by asking users what they need and which design options they prefer, are difficult to support on the basis of evidence.

Case in point: my students and I have gathered data indicating that early perceptions of usability seem to be heavily influenced by visual appeal or aesthetics and are not accurate indicators of how well people will be able to use an application. We have data showing that early dislike and poor performance with a system can give way, over time, to mastery and even a preference for the poor system over a better-designed alternative. And did you know that a recent German study of user-centered design processes revealed that greater user participation was related to less satisfaction with the resulting design? Add these up and you begin to see that unquestioning acceptance of users' views as the primary basis for designing information systems is simply naïve.

Does this mean we should exclude users? Of course not! They remain our best source of information on what needs to be built and how well a resource works for its intended audience. But much depends on what we ask of users and our abilities to interpret their responses. We must move beyond the unquestioning acceptance of all user data as a true and accurate representation of what needs to be built. User responses are subject to many forces, not all of them clearly recognized by the users themselves or the designers and evaluators who study them. User-centered design advocates have concentrated to date on gaining acceptance of their methods, and while progress has been made, this has been at the cost of strong research into the value and trust we can

place in certain types of user data. How many claims for usability, for example, are based on initial reactions to a system? Far too many, from my reading of the literature, and it now appears that in many circumstances, such initial reactions are poor indicators of actual use.

When we design our information spaces and invoke user-centered methods to test and evaluate them, how much consideration is given to the forces that drive the user response at that specific time? A quick usability test is certainly a good basis for gaining impressions but it can hardly tell us more. The most popular usability evaluation method in industry, the Heuristic evaluation method popularized by web guru Jakob Nielsen, is itself a limited, dare I say flawed, method that does not even employ users but rests on an evaluator inferring what a user would do and think (as if these were facts any evaluator really could infer). For me, this raises a whole other question of trust – that of trusting professionals who claim to serve the interests of users. It would be a good idea for us to take a critical look at the assumptions and analyses underlying the process of user-centered design.

Practicing user-centeredness requires more than asking users for their opinions or their time. It requires us to truly understand the complexities of user behavior and the forces that shape human actions and responses. William Horton remarked at IA 2000 that our designs often reflect our unconscious mistrust of users. It is about time that we raised the basis for such trust or mistrust to the conscious level. I suspect such an examination would cast doubt on many of the assumptions we have made about user-centeredness in this and related fields. Is the field of information architecture up to that challenge?