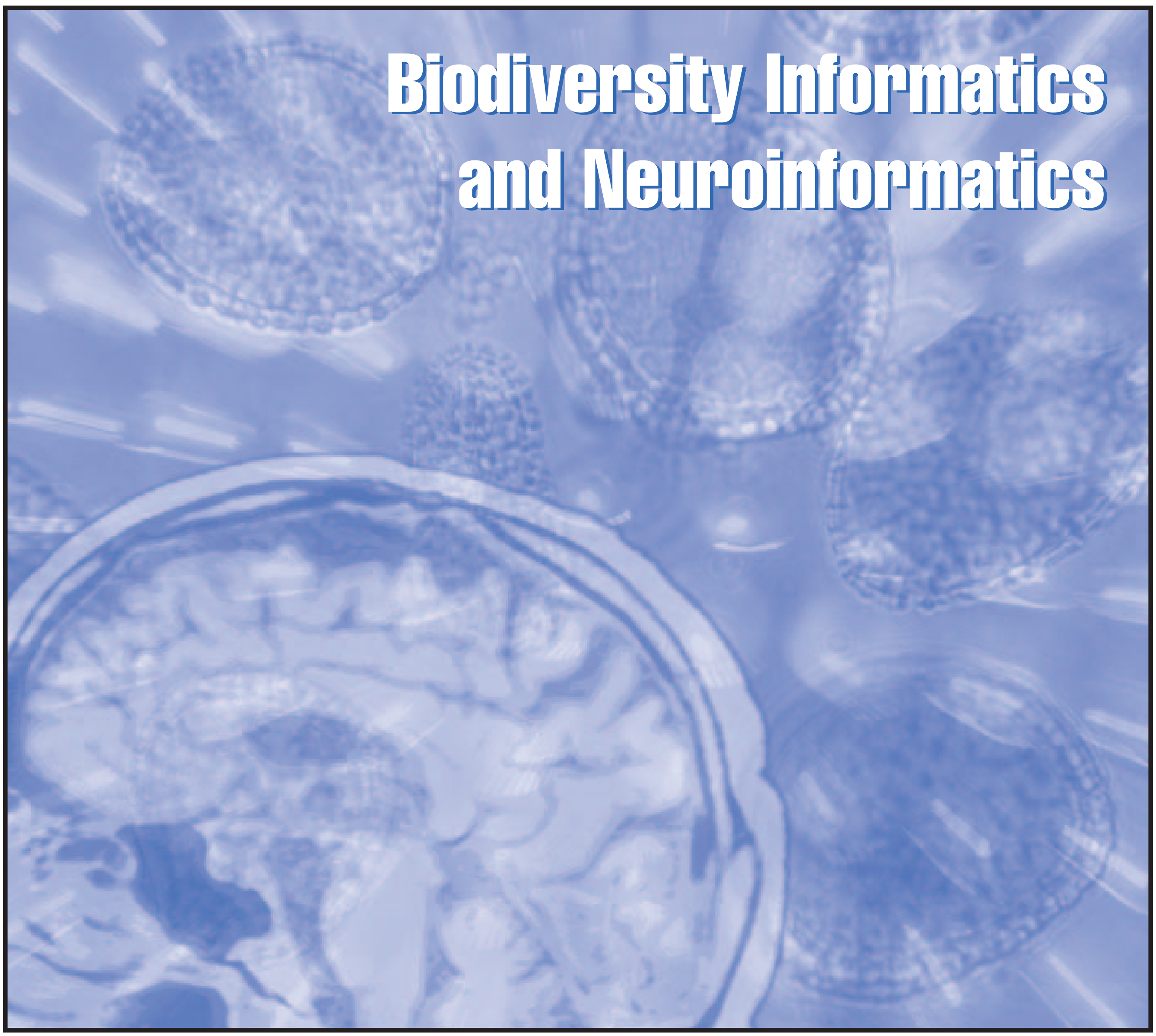
BULLETIN

of the American Society for Information Science and Technology *October/November 2003*

Volume 30, No. 1 ISSN: 0095-4403 CODEN:BASICR



OCTOBER/NOVEMBER 2003 VOLUME 30, NO. 1

BULLETIN

of the American Society for Information Science and Technology

SPECIAL SECTION

Biological Informatics

Feature

Columns

Departments

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; 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. Periodicals postage paid at Silver Spring, MD.

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.

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 galleys before publication.

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

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

I would like to use this, my last president's column, to reflect on some of ASIST's accomplishments this past year (note that I did not say "my" accomplishments – nothing here reflects my sole efforts). I also would like to suggest some directions for ASIST in the near future.

The past year flew by, as I knew it would. Nevertheless, a lot happened. I will not recap all the meetings, SIG and chapter activities, publications and other accomplishments that are well documented elsewhere; instead I would like to highlight some developments that were not so visible, but nonetheless are important for the continued health and viability of ASIST. This is only a sample; my apologies to others who could not be squeezed in this space.

Survey of Members

A Web-based survey was announced to all ASIST members individually in May. Given that ASIST does not have valid e-mail addresses for some members, the response of nearly 900 was gratifying. The results are being scrutinized closely and reported elsewhere.

2003 Annual Meeting Changes

Responding to members' suggestions, the 2003 Annual Meeting will start on Sunday and end Wednesday evening. Most of the awards will be presented at a luncheon on Tuesday; only the "secret" awards will be announced at the president's reception Wednesday evening. The member survey results included other suggestions related to Annual Meetings, which the Board of Directors will consider for the future.

New Model for Summits

Summits are a recent innovation for ASIST (the most notable and successful example has been the Information Architecture summit, held four years running now). Thanks to the joint efforts of the New England Chapter and SIG/Scientific and Technical Information Systems, we will have a different sort of summit in November on the MIT campus on the topic of digital archives in science and technology. This model for chapter and special interest group collaboration on a summit may be a powerful one for others to emulate, perhaps even for in a summit to be held in Europe.

International Members

Julian Warner, the international liaison to the Board of Directors, led a working group that submitted a thoughtful white paper of many issues related to international members. The analysis covered the value of international connectedness to both the international members themselves and to ASIST, new options to increase the number of international members, procedures to attract such members and strategies for moving these issues forward. The Board is giving these topics serious attention.

Mission and Goals for CE

Recognizing information professionals' needs for ongoing and cost-effective educational and learning opportunities, Liwen Vaughan chaired a subcommittee of the Education Committee that developed a fresh mission, goals, objectives and guidelines for ASIST continuing education. This document will spur increased activity in a wider range of technology use to make CE programs as accessible as possible.

Expansion of Placement and Job Services

Abby Goodrum headed a task force that is recommending a Web-based employment database (based on job board software) and augmentations to the Placement Center at Annual Meetings. Some of these changes will be evident at the 2003 Annual Meeting in Long Beach; others will be implemented in the coming year.

New Options for Institutional Memberships

Beverly Colby is leading a task force considering new institutional membership packages that will be mutually beneficial to ASIST and to organizations where many people in this field are working.

Greater Flexibility and Visibility of Special Interest Groups (SIGs)

The SIGs as a collective group, through SIG Cabinet, have embarked on a discussion of how to move from the traditional SIG structure to a more flexible and responsive one that meets members' diverse needs. SIGs represent the interests of their members – the communities of focus, which have a wide variety of informational and administrative requirements. The SIG Cabinet is also exploring expanded ways to communicate the interests and activities of their respective communities to members and prospective members. As always, they seek continual input on services offered.

Revamping the ASIST Web Site

As the young members might say, the current ASIST Web site is *so* 1990s. Thanks to a task force of the Board of Directors and volunteers from a Washington, DC, Web design team headed by Stacy Surla, we are getting a major site inventory and heuristic evaluation, including data from users' focus groups, which will result in a complete overhaul of the site.

ASIST Digital Library

An ASIST digital library, indexed by a new ASIST thesaurus, has been discussed for several years; it is likely that

(Continued on page 3)

From the Editor's Desktop

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

We begin a two-part special report in this issue on advances in biological informatics, focusing on the integrative challenges of biodiversity informatics and neuroinformatics. Guest Editor Bryan Heidorn has furnished six articles to which we have added a timely and pertinent announcement of the release of the *Biocomplexity Thesaurus*, an effort led by Jessica Milstead, winner of the ASIST 1999 Watson Davis for her many outstanding contributions to information science and to ASIST. I am very grateful to Bryan for pulling together such a fascinating collection of papers.

In our other feature article, "Copyright's Digital Dilemma Today," Lee Strickland provides a quick and highly informative tour of the litigation surrounding the MP3 file-sharing controversy. He issues an urgent warning about widespread potential liability for copyright infringement (yes, this means us!) with significant penalties attached, given the present aggressive enforcement actions of the Recording Industry Association of America. In the next issue he will review the most current legal developments regarding other kinds of e-copying, including the TEACH Act of 2002, to complete this survey.

The International Column features Julian Warner's views on recent decisions in the United Kingdom about the use of citation analysis in the Research Assessment Exercises that evaluate publicly funded university research. Julian is the international liaison to the ASIST Board of Directors.

Finally, out-going President Trudi Bellardo Hahn summarizes ASIST's accomplishments during her presidency. There are many significant changes underway from plans to revise the ASIST website to changes to the SIG administrative structures, the *Bulletin* and the Annual Meeting. This President's Page is definitely required reading.

I would like to thank Trudi for her varied and thoughtful reports and all our other contributors this year for their widely appreciated columns and articles. ASIST members provide most of our copy, and we rely on your efforts. I hope to see many of you in Long Beach. Please remember to pack your ideas for the *Bulletin* with your sunscreen. Westward Ho!

President's Page

(Continued from page 2)

it will become a reality soon. Until logistical and financial arrangements are finalized, however, it would be premature to announce it. I am excited, however, at the progress that is being made.

Revamping the ASIST *Bulletin*

For a few years we have been examining the idea of replacing the *Bulletin of the American Society for Information Science and Technology* with a new magazine focused specifically on applications of information science and technology, targeted at a wide audience. Given the current economics of publishing, it is unlikely that we will launch an entirely new publication in the near term. Nonetheless, the *Bulletin* editor, Irene Travis, and her Advisory Board will use the survey results to increase the appeal and usefulness of the current *Bulletin*.

Member Needs That Never Change

My year as president confirmed for me that in a long historical view of scientific and scholarly societies, some things never change. The recent survey confirmed these unchanging member needs:

1. Members need to learn about new developments in the field. We need to affiliate with others whose interests and work are similar or related to our own. Whether at face-to-face meetings, via electronic lists or reading publications, we recognize that connecting with others in our field inspires us, reinforces our shared paradigms, values and goals, and keeps us current.
2. Members need to network with others when we are seeking jobs or other professional opportunities or are seeking to hire people. We want also to connect with those who will commiserate about our common challenges and who will validate what we do.
3. Members need a sense of professional community and belonging to a recognized and significant field of study and research and development.
4. Members need to build skills, including such general skills as communication and leadership, as well as specific technical and research skills.
5. Members need recognition in the form of awards and other forms of credit, acknowledgment, appreciation and respect for our professional accomplishments.

My year as president has flown by, but it has been one of the most satisfying and gratifying experiences of my career. I regret that it is nearly over, but I am delighted to be passing the gavel to a strong, experienced and dedicated ASIST leader, Samantha Hastings.

Inside ASIST

LOOKING TO LONG BEACH

ASIST 2003 Annual Meeting – *Humanizing Information Technology: From Ideas to Bits and Back*

It's Annual Meeting Time

We're just a couple of weeks away from the start of the 2003 ASIST Annual Meeting, when ASIST members and other professionals will gather to address the theme, *Humanizing Information Technology: From Ideas to Bits and Back*. We'll meet in Long Beach, California, October 19-22.

The conference committee, chaired by **Marcia J. Bates**, UCLA, has planned a technical program that includes dozens of panel sessions, SIG programs, contributed papers and poster sessions throughout a four-day schedule. Bates was assisted in her work by committee members **Eileen Abels, Suresh Bhavnani, Michael Buckland, Donald Case, Chao-chen Chen, Andrew Dillon, Efthimis N. Efthimiadis, Raya Fidel, Jonathan Furner, Andrew Grove, Jenna Hartel, Sandra Hirsh, Joseph W. Janes, Don Kraft, Carol Kuhlthau, Marianne Nelson, Hope Olson, Carole Palmer, Jaime Pontigo, Dragomir Radev, Nancy Roderer, Victor Rosenberg, Linda Rudell-Betts, Bernie Sloan, Ross Todd, Irene Travis, Peiling Wang, Carolyn Watters, Judith Weedman** and **Barbara Wildemuth**.

Among the topics that will be addressed in the context of the conference theme are user-centered design, business and management informatics, virtual reference services, international information issues, organization of information, information policy and design for children. In addition, a number of popular continuing education and professional development workshops will be offered as pre-conference sessions. Details about specific meeting activities, including registration information for the conference and for professional development workshops, are available at the ASIST website.

Local Activities

Throughout the year, members of the **Los Angeles Chapter of ASIST (LACASIS)**, host chapter for the 2003 Annual Meeting, have given *Bulletin* readers a flavor of Long Beach and its environs through a series of articles in **Inside ASIST**. Now, as ASIST members make their final arrangements for their time in Southern California, we offer a few excerpts from those articles.

FROM FEBRUARY/MARCH 2003
by Bo-Gay Tong Salavador

Long Beach is not "just another beach city," by any means. The fifth largest city in California, its flagship attractions include the venerable Queen Mary, which houses a museum, a hotel and restaurants; the state-of-the-art Aquarium of the Pacific; the Museum of Latin American Art; and the Long Beach Museum of Art. A revitalized downtown showcases eclectic eateries, distinctive shopping and a vibrant nightlife offering jazz, blues and rock and roll. For the more athletically inclined, there are extensive walking and biking paths along the fabulous 5½ miles of shoreline, not to mention opportunities for swimming, surfing, sailing, jet skiing, fishing and even deep sea diving! And if you somehow overdose on the warmth of the sun and begin to miss the cold temps of winter, you can catch an Ice Dogs game at the Long Beach Arena. (Yes, Long Beach has its own professional ice hockey team).

If you yearn to visit attractions elsewhere in Southern California, driving distances are quite reasonable. Here are some approximate distances (in miles) from downtown Long Beach to a few popular destinations: Disneyland, 26 miles; Malibu Beach, 36; downtown Los Angeles, 25; Getty Center, 32; UCLA, 30; Hollywood, 32; Huntington Library and Gardens (San Marino), 28. If you prefer not to drive the freeways of Los Angeles, many destinations can be reached by public transportation. For instance, the Blue Line of the Metro Rail system begins just steps from the conference hotel and takes you to downtown Los Angeles. You can then transfer to the Red Line which transports you to many attractions in Hollywood as well as to Universal Studios.

FROM APRIL/MAY 2003
by Linda Heichman

Imagine a city on the beautiful California coast, a city alive with a bustling art scene – museums, galleries, monthly art walks, symphony, theater and opera. San Francisco? Think again. Los Angeles? Nope. San Diego? Nah. I know, Carmel. Good try, but no. It's Long Beach.

Not only does Long Beach boast an eclectic art scene, the city is home to world-class art museums, internationally renowned theater companies, its own symphony orchestra, opera company and master chorale. Museums include the Long Beach Museum of Art, housed in a Craftsman mansion overlooking the Pacific Ocean; the Museum of Latin American Art, the Western United States' premier museum of Spanish and Caribbean arts and culture; and the University Art

(Continued on page 3)

LOOKING TO LONG BEACH, continued

Museum at California State University, Long Beach (CSULB).

Performing arts abound in Long Beach. Choose from the Long Beach Symphony Orchestra, check out budding talent at California State University, Long Beach's Carpenter Performing Arts Center, Cal Rep, International City Theatre or Long Beach Playhouse.

Want more? Check these out.

Visual Arts

East Village Arts District – www.eastvillageartsdistrict.com

Long Beach Museum of Art – www.lbma.org

Museum of Latin American Art – www.molaa.com

University Art Museum – www.csulb.edu/org/uam/

Performing Arts

Cal Rep at Edison Theatre – www.calrep.org/

Carpenter Performing Arts Center – www.carpenterarts.org

International City Theatre – www.ictlongbeach.com/

Long Beach Playhouse – www.longbeachplayhouse.com/

Long Beach Symphony – www.lbso.org

From June/July 2003
by Claude Zachary and Linda McCann

The aviation industry has played an important part in the local economy, beginning with the Dominguez Air Meet in 1910, the first national air event ever held. The Long Beach Municipal Airport, the first in Southern California, was founded in 1924 by aviation pioneer Earl Daugherty. The real growth of the industry began with the opening of the Douglas Aircraft Company's facility in 1941 on a piece of land adjacent to the airport and the production of the C-47 (a military version of the DC-3) for the war effort. The facility has been in continuous operation ever since and is now run by the Boeing Company (www.boeing.com/commercial/facilities/longbeachsite.html).

The Historical Society of Long Beach is located just across the street from the Westin Hotel in the beautiful Breakers Building. Stop in if you can; they have a great collection (www.historicalsocietylb.org/index.html).

From August/September 2003
by Amy Wallace

So the annual conference theme caught your eye, and then you were sold by the ASIST Los Angeles Chapter (LACASIS) articles on Long Beach weather, arts, culture and history.

What if you want to venture beyond Long Beach? Once the conference is over you might want to take advantage of Southern California's wonderful weather and do some sightseeing. Everyone knows Southern California is famous for attractions like Disneyland, Universal Studios, San Diego Zoo, the Mighty Ducks, the Dodgers and the Angels – but there is so much more. What about a trip to Catalina Island, Venice Beach, the International Surfing Museum or the San Diego Super Computer Center? The website (www.lib.uci.edu/lacasis/about.html) includes points of interest in the Southern California area and approximate drive times.

And if all this information is not enough for you, LACASIS has members from all over Southern California who will be attending the Annual Meeting. Members will be happy to answer all of your questions about the local area.

See you in Long Beach!

News from ASIST Chapters

For its annual meeting, the **Northern Ohio Chapter of the American Society for Information Science and Technology (NORASIST)** planned a late August visit to the Strong Bindery, a custom bindery in Cleveland's Little Italy neighborhood. Bindery owner Ellen Strong, formerly owner of Coventry Books, has bound and repaired books for the Cleveland Orchestra, Cleveland Public Library, Cleveland State University, Kent State University, Ohio Archives, Western Reserve Historical Society and more than 2000 individuals who love books.

The **Southern Ohio Chapter (SOASIST)** and **Central Ohio Chapter (CO-ASIST)** jointly sponsored an August meeting featuring Thomas S. Blanton, director of the National Security Archive. Blanton spoke on the USA Patriot Act and its impact on libraries and civil liberties in general. He also discussed current and proposed exemptions and risks to the Freedom of Information Act.

SOASIST also notes that the latest edition of its award-winning newsletter, *SOASIST. …On the Move*, is online at www.soasist.org/enews/

The Los Angeles Chapter of the American Society for Information Science and Technology (LACASIS) held its annual fall workshop in mid-September. This year's session focused on "Database-Driven Websites," covering such topics as database-driven possibilities and unique applications; database technology in a nutshell; planning and building a database-driven website – the hits, hurdles and horizon; and using databases to serve up non-traditional resources. The keynote speaker was **Samantha Hastings**, ASIST president-elect and associate professor in the School of Library and Information Sciences at the University of North Texas.

The **New Jersey ASIST Chapter (NJASIST)** named **Christine L. Borgman**, professor, Department of Information Studies, UCLA, as the recipient of its 17th annual Distinguished Lectureship Award. The internationally known award, established in 1985, honors individuals who have made significant contributions to the field of information science.

Los Angeles Chapter Names Scholarship Winner

Joanne Orsatti from the University of New South Wales in Australia is the winner of the 12th Annual Margaret McKinley Memorial Student Scholarship Essay Competition, sponsored by the Los Angeles Chapter of ASIST (LACASIS). In addition, Lisa Nguyen, University of Hawaii at Manoa, and Weipang Yue, University of New South Wales, Australia, were named runners-up.

Orsatti has a bachelor degree with a double major in philosophy and science and technology studies and is currently finishing a master's degree in information science in the School of Information Systems, Technology and Management.

The McKinley Scholarship Essay Competition, established in 1992 in honor of the late Margaret McKinley, is offered to library and information students to encourage their participation in the activities of professional societies. The winner receives reimbursement funding up to $1,000 for registration, airfare and hotel expenses to attend the ASIST Annual Meeting and a one-year membership in ASIST. Two runners-up receive one-year memberships in ASIST.

News from ASIST SIGs

ASIST **Special Interest Group/International Information Issues (SIG/III)** has announced the top six winners of its fourth International Paper Contest. The following are the winners and their paper titles:

- Ifeanyichukwu F. Njoku, librarian, Federal College of Fisheries and Marine Technology, Victoria Island, Lagos, Nigeria, "The information needs and information seeking behavior of fishermen in Lagos State, Nigeria."
- Wathmanel Seneviratne, senior assistant librarian, University of Colombo, Sri Lanka, "Local information upload strategy in building up rural community information pages on the Internet: An approach to attend information needs of rural Sri Lanka."
- Tariq Ashraf, librarian, Institute for Integrated Learning in Management (IILM), Lodhi Road, New Delhi, India, "Information Technology and Public Policy: A Socio-Human Profile of Indian Digital Revolution."
- Partha Bhattacharya, executive consultant (Information & Documentation) FITT, IIT Delhi, Hauz Khas, New Delhi, India, "Advances in digital library initiatives: A developing country perspective."
- **Albert Arko-Cobbah**, Vista University, P.O. Box 380, Bloemfontein 9300, South Africa, "The role of libraries in student-centered learning: The case of students from the disadvantaged communities in South Africa."
- Joseph Kaduda, network manager, Centre of Geographical Medicine Research, KEMRI/Wellcome Trust Research Labs, Kilifi, Kenya, "A digital library in a rural malaria research centre in Africa: The Kenyan experience."

The winning papers were selected from among 51 papers received for this year's competition. Each winner will receive a two-year individual membership in ASIST. The first and second place winners of the competition will also receive a travel grant to attend the 2003 ASIST Annual Conference in Long Beach, California. These winning papers and other submitted papers for the competition will be considered for publication by the *Bulletin* and Elsevier's *International Information and Library Review*.

The following were members of this year's contest committee: **Sue O'Neill Johnson**, chair of SIG/III; **Yunfei Du**, chair of the International Paper Contest Chair; **Hong Xu**, University of Pittsburgh; **Nadia Caidi**, University of Toronto; **Liwen Vaughan**, University of Western Ontario; **Suzie Allard**, University of Tennessee; **Gretchen Whitney**, University of Tennessee; and **Nathalie Leroy**, United Nations.

News about ASIST Members

Charles W. Bailey, Jr., assistant dean for systems at the University of Houston, has been appointed assistant dean for digital library planning and development at the University of Houston Libraries. In 1989, Bailey established PACS-L, a mailing list about public-access computers in libraries, and The Public-Access Computer Systems Review, one of the first scholarly electronic journals published on the Internet. Bailey was profiled in *Library Journal's* Movers & Shakers 2003: The People Who Are Shaping the Future of Libraries.

Dick Hartley, head of the Department of Information and Communications at Manchester Metropolitan University, United Kingdom, has been awarded the title of Professor of Information Science.

Ronald Rousseau, Belgium, was international program chair of the 9th International Conference on Scientometrics and Informetrics, held in August in Beijing. He was also co-editor of the conference proceedings, published by Dalian University of Technology Press (ISBN 7-5611-2098-2). In addition to Rousseau, the following ASIST members gave oral presentations at the conference: **Henry Small**, USA; **Subbiah Arunachalam**, India; **Michael Nelson**, Canada; **Liming Liang**, China; **Leo Egghe**, Belgium; **Liwen Vaughan**, Canada; **Mike Thelwall**, UK; **Heting Chu**, USA; **Peter Ingwersen,** Denmark; and **Alastair Smith**, New Zealand.

Henry Small was elected as the new president of ISSI, the International Society for Scientometrics and Informetrics.

Feature

Copyright's Digital Dilemma Today: Fair Use or Unfair Constraints? Part 1: The Battle over File Sharing

by Lee S. Strickland

Editor's note: This article has been split into two parts. The first covers the legal controversy over file sharing. The second, to be published in the December/January 2004 issue of the *Bulletin*, covers other critical developments in e-copying.

Lee S. Strickland, J.D., is visiting professor in the College of Information Studies, University of Maryland; e-mail: LSTRICKL@deans.umd.edu

In previous issues we surveyed some of the critical legal and policy developments in the world of records and information management (*Bulletin*, June/July and August/September 2003) – a key concern for every business whether directly focused on the task of homeland security or more generally contributing to our national economic interests. Today we continue our survey of cutting-edge developments of interest to information professionals with a focus on the most current intellectual property issues – the ultimate business asset in our information age. As with our previous articles, we also provide recommendations for management steps as well as additional research and sources for maintaining currency in this rapidly developing arena. *Especially critical for every business and every individual is our discussion of potential and almost certain liability for infringement given the proliferation of peer-to-peer (P2P) network connections in the corporate environment and among home users.*

The Contest Over Intellectual Property Rights

Twenty-five years ago, perhaps even 10 years ago, intellectual property law in general and copyright law in particular were of importance to information professionals but were relatively static from legal and policy perspectives. But more recently, our information age and information economy have propelled this subject to extreme importance and exponential change – in large part because technology has vastly altered the medium of intellectual property. The result has been efforts by millions to gain access to intellectual property on their terms – such as unauthorized electronic sharing of copyrighted music – and, in response, by business to protect their content by increasingly aggressive tactics.

In response to these drivers, the Computer Science and Telecommunications Board (CTSB) of the National Academies of Science issued its seminal report in 2000 – *The Digital Dilemma: Intellectual Property in the Information Age.* The study was predicated on the truism that the very technologies underlying the information age and providing extraordinary levels of access to information were also greatly enhancing infringement. As a result, the CTSB concluded that our current intellectual property laws – developed in the physical world of paper and tangible goods – simply did not work well in the digital world. Although the CTSB offered many thoughtful proposals, the approach of content providers has changed little. Indeed positions appear to have hardened with industry demanding greater control in a multiplicity of ways and users fearing that traditional fair use concepts will be eroded if not eliminated. And thrust into this contest were the courts and the Congress as they have struggled to address this tsunami of transformation.

As information professionals, we have a primary responsibility for understanding these rapid developments and shaping the underlying policy debates. For example, how has the electronic environment changed the basics of established copyright law? Should it? How is the Digital Millennium Copyright Act (DMCA) being used – as expected by Congress or in unforeseen ways? What are the rules for public domain today for published and unpublished materials since the Supreme Court has now upheld the Copyright Term Extension Act. And what does the TEACH Act mean for me? These are some of the many questions we will explore.

E-Copying — New Technology Does Not Change Established Law

Where previously there were photocopy machines and singleton copies there is now electronic reproduction with unlimited copies. But from a legal perspective, electronic copying is simply traditional infringement with new tools, and a number of litigations have confirmed that cyberspace does not change traditional law. The most visible case evidencing this truism was *A&M Records, et al. v. Napster*, filed in December of 1999 in federal court in San Francisco by a number of record companies against the Internet company that provided a frequently used search tool for locating MP3 music files on the PCs of other, individual Internet users. As we know the defendant did not itself provide or host music files and the suit was therefore based on the allegation that its service facilitated copyright infringement and that it was therefore liable under the established doctrine of contributory infringement.

Initially, on July 26, 2000, following detailed hearings and consideration, the district court granted a preliminary injunction on the grounds that consumers who use Napster service and software to exchange sound files containing copyrighted musical recordings are engaged in direct copyright infringement and that Napster itself is liable for contributory infringement as well as vicarious infringement. However, two days later, the U.S. Court of Appeals for the 9th Circuit granted a stay nine hours before Napster was to shut down – finding that Napster's appeal *"raised substantial questions (on) both the merits and the form of the injunction"* and setting a schedule for expedited consideration. That followed in October 2000 with, generally considered, critical questions posed to the record industry lawyers. At the same time, the U.S. Copyright Office filed an amicus brief arguing, in part, that the Audio Home Recording Act was not a relevant defense.

But predictions of decisions based on factors such as critical questioning by appellate judges are notoriously unreliable – on February 12, 2001, the 9th Circuit affirmed the decision of the trial court framing the following issues and holdings and providing us important guidance for the future.

Is there direct infringement by the individual users of Napster or do they have a valid defense such as fair use? The answer was direct infringement given the four statutory fair use factors. First, as to "purpose and character of the use," the copying was not transformative and was commercial in nature given the scope and objective (i.e., "to save the expense of purchasing the authorized copies"). As to "nature of the use," the works are creative and thus closest to core protection and furthest from a finding of fair use. As to "portion of use," the works were copies *en toto* and in massive or "wholesale" quantities. And as to "effect of use on market," there was material impact given expert testimony; and, in any event, a lack of harm to an established market is not the only issue since harm can come from preventing plaintiffs' development of a secondary electronic market for their goods.

If the individual users are liable, then is Napster liable under a theory of contributory infringement or vicarious infringement? The answer was "yes" to both questions. First, contributory infringement requires a finding of "*…knowledge of the infringing activity and personal conduct that encourages or assists the infringement."* Here, Napster had both actual and constructive knowledge that its users exchanged copyrighted music. Moreover, the *Sony* decision is of no use given Napster's actual specific knowledge; the Sony court had refused only to impute knowledge given the potential of the equipment to infringe. Second, vicarious infringement requires a finding of *"…the right and ability to supervise the infringing activity and also a direct financial interest in such activities."* Here, Napster has the ability and authority to block infringers but only a limited ability to identify copyrighted information – it does not read the contents of the MP3 files and can identify copyrighted material only by the "titles" as listed by its users. It also has a financial interest since the infringing material "acts as a draw." Given the limits on Napster's ability to identify infringing material – and the equal ability of the copyright holders to do so – the Court found that this ground required modification of the injunction to require identification by the copyright holders but then to require resolute remedial action by Napster.

Does the Audio Home Recording Act apply to this case? The answer was "no" inasmuch as computer hard drives are not defined in the Act as *"digital audio recording devices"* and the MP3 recordings on computer hard drives are not defined as *"digital music recordings."*

Does "sampling" and "space (or time) shifting" by individual users constitute fair use? Here the court found these arguments not only unavailing but inapplicable. Specifically, "sampling" is in essence "promotional downloads" which is a tightly regulated activity by the plaintiffs generating substantial royalties from the process; hence, the free sampling would not constitute fair use. And, the shifting argument (relying on the *Rio* and *Sony* cases) was similarly inapplicable since this case did not involve an individual user moving a file legally owned or acquired from one place to another but rather involved massive distribution of copyrighted material to the general public.

Lastly, are the affirmative defenses of waiver, implied license and copyright misuse applicable? Again, the answer was "no." Just because the music industry has promoted MP3 technology for legal purposes does not mean that it has waived its intellectual property interests, including its authority to exercise exclusive control over the creation and distribution of MP3 files of its music.

In conclusion, the 9th Circuit found infringement on the part of Napster but remanded only to require the plaintiffs to provide notice to Napster of copyrighted works and files prefatory to Napster's burden and duty of disabling access to the offending content. And, without question, the Court reiterated that Napster bore the burden of policing its system within the technical and practical limits of that system. In relatively quick succession, an injunction followed in March 2001 requiring both parties to take action to identify and remove infringing materials, the company voluntarily took down its servers to attempt compliance and the company failed in its efforts to re-invent itself as a subscription-based, pay-for-download service despite an abortive partnership and later sale to German media giant Bertelsmann AG.

The Stage Shifts – Decentralization

With the demise of Napster, a bevy of new music-trading software products emerged (e.g., Aimster, Kazaa and Morpheus) that were based on a true peer-to-peer architecture with no central control and no central index. Here, when users download the software, they become members of this network and can share files directly with other members – whether those files are innocent family photographs or copyrighted music. And, as might be expected, the recording and movie industry lost no time in bringing federal lawsuits against Aimster (later renamed Madster) in Chicago and Morpheus (and others) in Los Angeles.

> ***A technology note:*** The P2P network is not truly index-free – rather the software distributing company (and most visible litigation target) does not maintain any index or specific knowledge of traded content. In fact, the index responsibility is assigned to certain users. More specifically, any Kazaa user having a broadband connection will be designated a "supernode" and will thus automatically maintain a list of some of the files (and IP addresses) made available by other Kazaa users, usually on the same Internet Service Provider (ISP) network. When a given Kazaa user initiates a search, the software first queries the nearest supernode and it provides immediate results; that first supernode then refers the search to other supernodes (and so forth) as necessary. The actual file sharing takes place directly between the requesting PC and the target PC – not through the supernode and not the software company. As is evident, every member of the Kazaa network that functions as a supernode quite arguably has the same legal liability as Napster for contributory infringement. And, of course, every individual user is similarly and potentially liable for direct or contributory infringement depending on whether that user is downloading or simply making available a copyrighted work.

But there were legal problems based on the technology at issue – couldn't these services more clearly claim the *Sony* defense since they had no knowledge of and control over individual trading? Remember that in *Sony*, the Supreme Court held that the sale of VCRs did not constitute contributory infringement even though Sony knew in general that the machines could be used and perhaps were being used to infringe copyright. This was because the VCRs were capable of both infringing and "substantial noninfringing uses" and the fact of generic or constructive knowledge of infringing activity was insufficient to give rise to liability given only the sale of the devices. The clear holding was that the mere sale of items in commerce does not constitute a civil wrong if the items are capable of appropriate uses. Indeed the courts have often used the analogy of "sexy lingerie and prostitution" to explain the logic underlying this decision.

In any event, the cases moved forward to very recent fruition. First came the A*imster* litigation by the Recording Industry Association of America (RIAA) and the P2P service was shut down in short order by the U.S. District Court in Chicago on a preliminary injunction pending trial. And that decision was promptly upheld in July by the U.S. Court of Appeals for the 7th Circuit. Why was this so clear despite the *Sony* defense? Quite simply, Aimster suffered from their bravura and was quite unable to demonstrate any legal issues of their service as stated by the appeals court: *"Far from doing anything to discourage repeat infringers of the plaintiffs' copyrights, Aimster invited them to do so, showed them how they could do so with ease using its system and by teaching its users how to encrypt their unlawful distribution of copyrighted materials, disabled itself from doing anything to prevent infringement."*

But between the two decisions in *Aimster* came the *Grokster/Morpheus* decision based on summary judgment arguments in December 2002. In a decision that surprised many, given the string of industry victories to date, the trial court on April 25, 2003, ruled in favor of two of the defendants – Grokster and StreamCast Networks (distributor of Morpheus) – finding no direct, contributory or vicarious liability based on the defendants' lack of control of individual users and the fact that the software had valid non-infringing uses. In other words, the arguments based on the 1984 U.S. Supreme Court decision involving the Sony Betamax had won the day.

How could this be? How could the Sony defense win and lose seemingly identical cases? Quite simply, it was a matter of a better defense and better proof. Here, the defendants were able to establish that there were substantial noninfringing uses for their software: distributing movie trailers, free songs and other non-copyrighted works including Shakespearean plays, government documents and other public domain materials.

But this is not to suggest that the story ends here. The RIAA responded testily that *"[b]usinesses that intentionally facilitate massive piracy should not be able to evade responsibility for their actions"* and has appealed the decision, although success in the Court of Appeals is doubtful given the thoughtful, fact-based nature of the decision. What is significant in the strategic sense is the observation of the trial court that it is not *"blind to the possibility that Defendants may have intentionally structured their businesses to avoid secondary liability for copyright infringement, while benefiting financially from the illicit draw of their wares…* [and that]*…additional legislative guidance may be well-counseled."* As we shall see, it appears certain that the music industry will continue the fight at both the individual and the Congressional level.

Resource notes: The primary decisions concerning technology and the liability for direct or contributory infringement are:

- *Sony Corp. of America v. Universal City Studios, Inc.*, 464 U.S. 417 (1984).
- *A&M Records, et al., v. Napster*, 239 F.3d 1004 (9th Cir. 2001)
- *MGM Studies v. Grokster*, 2003 U.S. Dist. LEXIS 6994 (C.D. Cal. 25 April 2003)
- *In re Aimster Copyright Infringement*, 2003 U.S. App. LEXIS 13229 (7th Cir. 30 June 2003)

Decentralized Sharing and Innocent Service Providers

Given the demise of Napster and the rise of decentralized sharing, industry attention has also focused on service providers that provide the connectivity for P2P users. One favorable feature of the DMCA (section 512(c)(1)) was the limitations of liability for service providers otherwise known as the "safe harbor" provisions – a significant step given past cases that had held service providers liable for vicarious infringement under the 1976 Copyright Act. In general it provides a defense from copyright liability for typical operations if the provider (1) has rules to prohibit infringement and to terminate repeat offenders, (2) has no knowledge of infringement, (3) does not interfere with any technical schemes to protect copyright data and (4) acts expeditiously to remedy infringement upon notice. This protection however is dependent upon registration with the U.S. Copyright Office (USCO) and the adoption of a set of policies that regulate user activities (e.g., acceptable use policy) as well as specific "take down" and "put back" policies that guide activities when a notice of infringement is received including required notice to the individual. It is important to note that the definition of service provider is very broad – any entity providing online services or network access – and clearly could encompass every public or private business today, including, of course, libraries and educational institutions.

The U.S. Naval Academy (USNA) is case in point. Last November, in response to RIAA demands, authorities seized nearly 100 student computers suspected of containing copyright-protected content. This issue is bedeviling college administrators across the country where students have the benefit of broadband Internet connections and where the copyright industry is well aware and well focused on this threat. Critical to continued immunity is the adoption of computer policies that prohibit students from utilizing the institutional network to access and transmit copyrighted content and to penalize offenders (e.g., immediate termination of use).

Three final factoids of importance are the extent of the demands and the research that must be accomplished before such demands are made by copyright holders and resulting action by service providers. First, it appears that the USNA action was initiated as part of a mass effort by the recording and movie industry to over 2,000 colleges and universities – an effort that continues today and has an increasing focus. Second, the factual predicate to such demands is very low – perhaps good news for content providers and service providers but bad news for alleged offenders. According to a very recent decision in *Motion Picture Association of America (MPAA) vs. InternetMovies.com* by the U.S. District Court in Hawaii, the DMCA requires no investigation before asserting a takedown notice even if the claim is unresearched and/or inaccurate. And third, the DMCA immunities provision also precludes a lawsuit against the service provider by the individual alleged to be responsible for the infringing material – a not insignificant point in this litigious age.

A practice note: The fee to register for the DMCA immunity is currently $30.00 and the forms are available online at the following addresses:

- www.copyright.gov/onlinesp/agent.pdf (for initial designation)
- www.copyright.gov/onlinesp/agenta.pdf (for amendments).

The Battle Shifts to Individuals

Little more than a year ago, §512 of the Digital Millennium Copyright Act (DMCA) was a little-noted provision but harbored potentially great power: it allowed for automatic issuance of subpoenas in DMCA cases without any judicial consideration. Today, the RIAA and Verizon have made this section a feared weapon in the copyright wars.

Here is the story. After crushing the visible, big infringers such as Napster, it was expected that the RIAA would begin to move against individual alleged violators – especially, as we considered previously, technology moved toward P2P. But how could the music industry find allegedly infringing material on individual's machines? The answer is part technical, part guesswork and part legal. The industry finds content it deems infringing the same way that you find music that you want to copy (i.e., internet-shared folders on your machine). The industry identifies the content as copyright-protected by guesswork (i.e., the designated file name appears "close enough" to their copyright catalog). And the industry identifies the name and address of individuals offering or downloading that content through a §512 subpoena to their ISP. The legal and policy issue? These subpoenas are issued by the music company itself, without any pending litigation, without any showing of necessity or violation of law, without any judicial review and without any right of the target to contest the matter.

In short order, the RIAA issued numerous subpoenas and much to the credit of Verizon (in its role as an ISP), the company refused to release the demanded customer data. RIAA thereafter filed a motion in the U.S. District Court for the District of Columbia to enforce compliance in October of 2002. Unfortunately, individual rights were the loser: first, on January 22, 2003, the court rejected Verizon's statutory challenges based on the fact that the RIAA's subpoena related to material transmitted over Verizon's network (rather than stored on it) and thus fell outside the scope of the subpoena power authorized by §512(h). And second on April 24, 2003, the court also rejected Verizon's constitutional claims based on the argument that the subpoena power violates Article III of the Constitution (because it authorizes federal courts to issue binding process in the absence of a pending case or controversy) as well as the First Amendment rights of Internet users. A week later, the D.C. Circuit Court of Appeals refused a stay, thus requiring immediate compliance by Verizon of the release the names and addresses.

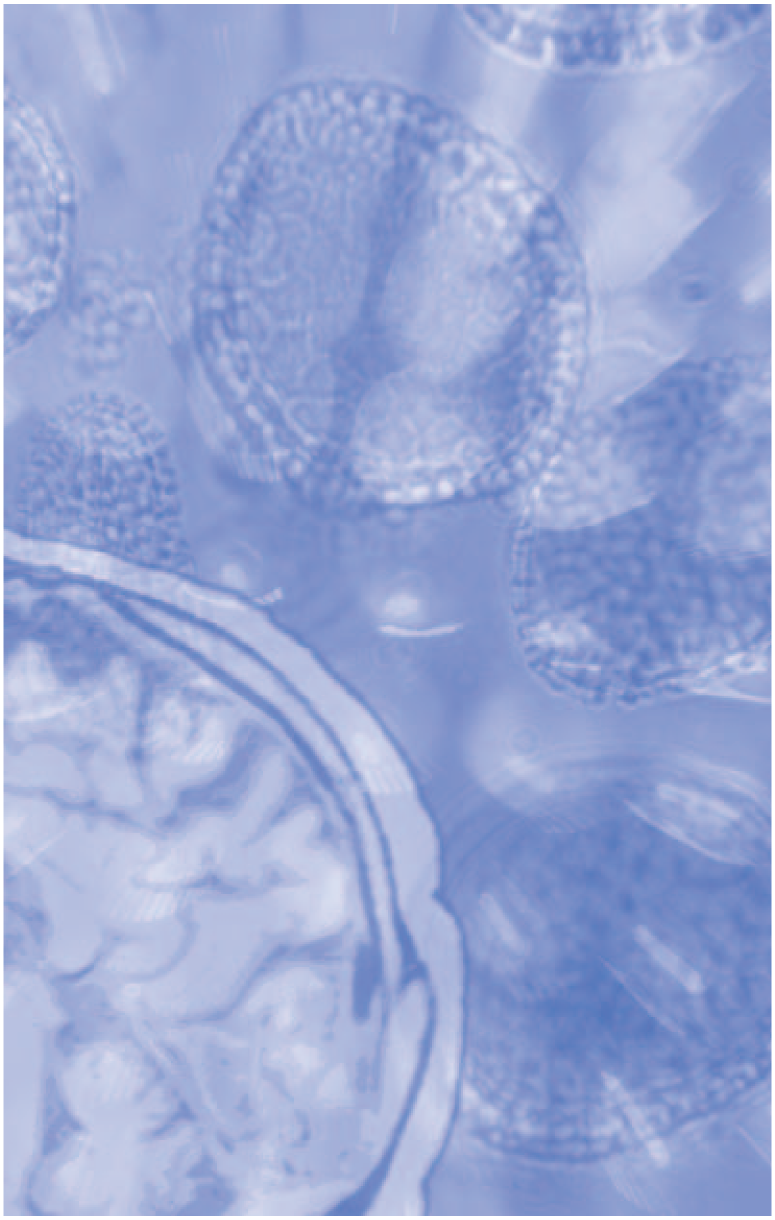
A Risk Assessment for Networks and Individuals

The risk posed by the recording industry to individuals trading copyrighted content – as well as corporate entities that provide connectivity to employees, students and others – is substantial and real. It is substantial in terms of civil and criminal penalties. Civil remedies for copyright infringement may include *actual damages* or *statutory damages* – an alternative form of recovery since actual damages may be difficult to prove or limited in dollar amount in typical cases. Statutory damages are specified in three categories for each single infringement of a single work: *normal* with damages from $750 to $30,000; *willful* with damages increasing to $150,000; and *innocent* with a court allowed to reduce damages to an amount not less than $200. Clearly the stage is being set to obviate any claim of innocent infringement and impose willful penalties. Moreover, criminal penalties for individuals are also a real risk especially given the passage of the No Electronic Theft (NET) Act, an amendment in 1997 that removed the personal financial gain or commercial advantage element and provides for penalties that include up to five years in prison and/or $250,000 fines and a statute of limitations extended from three to five years.

And it is real in terms of the extent of infringement and the clear actions underway by the industry. Although a substantive appeal of the Verizon litigation was to be heard in September 2003 by the Court of Appeals, the wheels of litigation against individuals began almost immediately – on June 26 the RIAA announced publicly that they were *"preparing a wave of civil lawsuits against people who use music-trading software."* Targeting would include the public directories of peer-to-peer providers and according to privacy advocates at the Electronic Privacy Information Center (EPIC), some 57 million Americans could be liable to these *"dinosaurs [that] have completely lost touch with reality."* As of late July and reported by the *Washington Post*, over 1,000 such subpoenas had been issued with a rate of at least 75 additional per day.

The corporate environment – from private businesses to educational institutions – is no different. According to a very recent survey by AssetMatrix, a Canadian network monitoring company, as reported by CNET News.com, the installation of P2P clients in corporate environments is more than common – it is pervasive. While this would be expected in education, it is somewhat of a surprise in private industry. The study considered over 500 companies of varying size and found that every company of 500 or more employees had at least one installation of Kazaa or similar sharing software. The president of AssetMatrix was cogent in his observation: *"Corporations are frantic about how to rein in some control over this. Like with software licenses, most companies want to be on the right side of the law. The challenge is how they do that."* As with the effort against educational institutions, the focus of copyright holders and the legal liability of corporations are certain – last year, the RIAA settled a copyright claim against Integrated Information Systems for $1 million. At that time the senior vice president for the RIAA noted: *"This sends a clear message that there are consequences if companies allow their resources to further copyright infringement."* In the most recent action (July 2), the RIAA served a DMCA subpoena on DePaul University in Chicago. This action presents the very practical issue of what quantity of offending material will trigger RIAA action. While the RIAA has previously suggested that they will target only *"substantial collections,"* the DePaul action appears to involve a user with only a handful of potentially infringing files.

In the next issue we will continue our survey of current intellectual property developments.



Biological Informatics: A Comparison of Biodiversity Informatics and Neuroinformatics

by Bryan Heidorn, Guest Editor

This special section of the *Bulletin of the American Society for Information Science and Technology* is devoted to two branches of biological informatics that warrant special attention from the information science community: biodiversity informatics and neuroinformatics. Biologists are turning to information technology to produce critically needed efficiencies in their work, a hope conveyed by E. O. Wilson (see box). Unfortunately, most of the information science community is unfamiliar with either of these fields of study so definitions and explanations are in order.

The relationship between *bioinformatics* and the term that we have chosen to use here, *biological informatics*, is not as subtle as it might seem. Over the past 10 years *bioinformatics* has come to mean "information on molecular biology" and in particular gene and protein sequences. This use of the term in the popular press, associated with the great progress and successes in that field, has served to cement this definition into the psyches of the general population and scientists alike. As a result, a new term (*biological informatics*) to cover the science of information about all levels of biologi-

P. Bryan Heidorn is assistant professor, GSLIS, University of Illinois, Urbana-Champaign. He can be reached by mail at 501 E. Daniel St., Champaign, IL 61820; by phone: 217-644-7792; or by e-mail: heidorn@alexia.lis.uiuc.edu

> "New electronic technology, increasing exponentially in power, is trimming the cost and time required for taxonomic description and data analysis. It promises to speed traditional systematics by two orders of magnitude. What is lacking and needed now is a concerted effort, comparable to the Human Genome Project (HGP), to complete a global biodiversity survey – pole to pole, whales to bacteria, and in a reasonably short period of time." (Wilson, E.O. (2000). A global biodiversity map. *Science 289*, p. 2279.)

cal analysis was in order. Health informatics (medical informatics), neuroinformatics, biodiversity informatics and biomolecular informatics (bioinformatics) all fall under this broader concept. Because of the length of this special section dealing with biological informatics, it will be split between this issue and the next one.

Both biodiversity informatics (BDI) and neuroinformatics (NI) are brought together into one category because, while they differ broadly in other aspects of their social, political and historical context, it has become clear that progress in both fields is dependent on advances in information gathering and access. Both have reached the point of becoming "integrative sciences." In this issue of the *Bulletin*, the paper "Toward Integrative Science," by Melissa Cragin, addresses how expansion and fragmentation of knowledge has led to the opportunity and need for integration. This integration and sharing of data is a new concept for many of the scientists at the center of this process. Geoffrey Levin and Melissa Cragin address the role of data gathering and integration across laboratories and institutions in "The Role of Information Science in Gathering Biodiversity and Neuroscience Data."

Progress has been made in biological informatics, including the formation of the Global Biodiversity Information Facility (www.gbif.org). Meredith Lane discusses this facility and related global projects. Neuroscience has a similar push for integration of information to answer new questions that could not be addressed using the traditional paradigm of disposable data.

Both BDI and NI are marked with a critical need for advancement, an abundance of new research and new sources of information, but a lack of a cultural and technical infrastructure to share and use this abundance of information. As is true in many fields, electronic media and the Internet are changing the nature of publishing in BDI and NI. Publishing projects now can include hundreds of authors spread across the globe. In endeavors like *The Flora of North America, The Flora of Australia* and *The Flora of China* the goals are frequently to document all species in a geographic region or all species in a particular group such as the grasses. In either case the species can number in the tens of thousands. Information science needs to develop information management methods to help scientists coordinate their efforts to document all species on earth. These same forces of scale and integration are breaking down the culture of "one lab, one data collection." Bryan Heidorn will address the promise and perils of this change in "Publishing Digital Flora and Fauna" in the December/January 2004 issue while Gwen Williams, in the paper "Intellectual Property and Biological Knowledge," will discuss how the shift to electronic delivery has made issues of intellectual property in biological science more complicated.

Assuming that we can solve all of the above issues, we still have a major problem with the physical communication of biological information. Projects springing up around the world are being designed to use existing information retrieval standards to record, encode and deliver biological information. Chulin Meng will cover some of the emerging standards with familiar sounding names but new twists, such as "Darwin Core" (tsadev.speciesanalyst.net), Z39.50 (ZBIG) and Open Archives Initiative style treatment of natural history style holdings in "Biological Information Standards" (December/January 2004).

We are still just beginning a biological science revolution. Global environmental change, extinction, alien species invasion, over fishing and many other natural and man-made phenomena are making it essential that we understand our world's biodiversity so that we can preserve the pieces that we need for survival. Advances in neuroscience are leading the way to our understanding normal brain function and treatment of human diseases of the nervous system. Information science and technology will play a central role in this revolution.

Toward Integrative Science: Organizing Biodiversity and Neuroscience Data

by Melissa H. Cragin

Melissa H. Cragin is a doctoral student in the Graduate School of Library and Information Science at the University of Illinois, Urbana-Champaign, 501 E. Daniel St., Champaign, IL 61820. She can be reached at 217-244-0866 or by e-mail: cragin@uiuc.edu.

The 1999 OECD report on biological informatics brought the biodiversity and neuroscience fields together into a single framework to describe needs and barriers to global information access. These two fields share a common orientation, focused on finding and identifying objects, naming new organisms and structures, and discovering and describing the relationships between them. The need for integrative science in the biodiversity and neuroscience fields has real consequences, and many important questions will only be addressed by integrating data from many and varied sources. Often such integration will mean connecting data derived from different levels of analysis, different species, even different fields. This paper presents some of the challenges researchers face in bringing data sets together for retrieval and use, including the creation of data collections, collection interoperability, and access and use.

Both fields are entering a stage of development where new discoveries will be made by learning how systems and subsystems behave. Without an integrative approach to biodiversity data, political and economic policy decisions are made with very little information about the big picture. That is, understanding biodiversity on a global scale will require research beyond the organismal level. It will require an understanding of the land itself, as well as weather and other phenomena. Data collected for very different purposes, like rainfall data, will need to be included in retrieval systems for analysis along with species data. Similarly, in neuroscience, advances are made daily that support our understanding of individual cells and subsystems of the brain, but developing a complete picture of the brain will require methods that can support evaluation of processes which are themselves integrative. To understand the processes of normal nervous system function and disease states, scientists will need to synthesize findings from different levels of study and determine their interactions and relationships.

There are also key differences between biodiversity and neuroscience, such as the structures of the fields, the management of data and the traditions of interaction with other (or sub) disciplines. Biologists and systematists, for instance, recognize the need for communication and collaboration with other fields such as ecology. They have come to value wide access to biodiversity databases that will connect with data sets created by a variety of owners. Neuroscientists, however, do not have a tradition of sharing raw data (Chicurel, 2000; Koslow, 2000); there are no projects to digitize legacy data for future use. Also, although neuroscience may be more interdisciplinary in practice than the field of biodiversity, it is a very "fractured" science in which subspecialties have developed along several dimensions (Toga, 2002). Specialties such as neuroimaging, electrophysiology, morphometrics and computational modeling each shape practices resulting in an array of preferred instrumentation and methods.

In order to take advantage of informatics developments scientists will need to change some of their research practices. The most prominent concerns identified in the literature of both fields are about data quality and trust, data sharing, communications with external groups or disciplines and rights management. Just as important but less discussed are the ways that researchers think about future uses of their data, including the following:

- How does their data relate to research outside of their specialty?
- Are there possible uses for their data in different kinds of studies?
- What information will be necessary to accompany a data set so that it can be re-used?

Gathering Data Together into Collections

Successful integrative science will require more open scientific practices in which scientists are cognizant of possible use of their data in the future, and institutions value and reward activities beyond traditional publication (Amari, et al, 2002). For the scientists such practices will include documenting the processes of data production and submitting that data and documentation to a repository or building a new shared database. For institutions this will require recognizing these new kinds of contributions in tenure and promotion processes. For long-term use databases and repositories will need curatorial management and archiving expertise to preserve the contents. Considerations of sustainability should be a mandatory part of these large-scale projects, and this will require addressing issues like funding, control and access. First, however, there needs to be consideration of what data should be included in shared databases and repositories.

Legacy Data. Legacy data are those that result from research conducted in the past that are not yet stored digitally. In biodiversity these data include museum data, conservation data and data found in individual labs. Some biodiversity data are simple and relatively homogenous. Such data generally consist of label data from biological specimens, including the scientific name of the specimen, where and when the specimen was found and who found it. However, as noted elsewhere, knowledge about the species that inhabit the planet will depend on the integration of data from other fields. If label data from many collections over a 100-year period are pooled and combined with forest fire data, it would be possible to evaluate the role of fire in species distribution, which could help set burn and forest clearing policy.

Integrating Scales of Analysis. A primary goal of biological informatics is to shift the focus from research on discrete pieces of a biological puzzle to whole sections of the puzzle. Sometimes this work involves the integration and analysis of data at different scales. Some of the most vivid examples of this can be seen in the work conducted at the National Center for Microscopy and Imaging Research (NCMIR) where informatics researchers are designing tools for scientists to literally see microscopic structures in degrees. Scientists use these tools to locate and render images of very small parts that actually vary in size, from spine to dendrite, synapse to neuron to cell cluster. Scientists are creating images of materials that scale with a factor of ten or more. The complexity of merging such different units of analysis is uniquely challenging, and will require new visualization and classification tools.

> "The Human Brain Project will shape, into the next century, an important frontier of science and technology…. The long-term goal of this initiative is to integrate brain and behavioral research with informatics research and development, giving scientists, students and clinicians intelligent access to the full range of information about the brain. A newly integrated view of the brain will facilitate the generation of hypotheses that address broad issues that are significant to the overall understanding of human health and disease." — Koslow, S.H., & Huerta, M.F. (Eds.). (1997). *Neuroinformatics: An overview of the Human Brian Project.* Mahwah: NJ: Lawrence Erlbaum Associates, p. 9-10.

Developing Collections of Tools. The collection of analysis and visualization tools is already an important aspect of biological data systems. For access to the tools at the NCMIR, scientists may either come to the center or utilize telecomputing. This serves the research interests of both the scientists, who may not otherwise have access to these tools, and the center that is fostering collaborative research. Neuroscience has traditionally been a local activity – data systems and tools have been developed by or for the specific labs in which they are used. This practice continues, though now tools are also developed by those building the common databases and repositories with the intention that anyone can use the tools when they access the data. However, some scientists still want to be able to use their own locally developed tools on this shared data. To address this desire, some initiatives such as the Biomedical Informatics Research Network are collecting locally developed tools and making them centrally available to users, while protecting the "ownership" rights of the developers. Still, proprietary issues and concerns about standardization persist in the field, and these issues point to the kinds of theoretical problems described by Bowker (2001a, 2001b) on the integration of biodiversity data.

Accumulating and Mobilizing Literatures. Neuroscience is a very broad field, with subdisciplines spanning anatomy, molecular biology, neurochemistry, electrophysiology, neuroimaging and behavior. As noted, research has become highly specialized, and increasing fragmentation results in a large

Although computer storage devices continue to increase in capacity, they will not be able to hold everything. Further, disparate sets of data require different kinds of organization for management and querying.

scatter factor. Researchers acknowledge that they have difficulty “keeping up,” even in their own areas of expertise. An early goal of the Human Brain Project was the development of an atlas of the human brain “to be used as a framework to index and link to databases for other types of brain data, including those not necessarily captured by neuroimaging (e.g., histological data, models of function of particular brain structures, directories of scientists studying certain brain structures, and bibliographic data)” (Huerta & Koslow, 1996, p. S 5). Although such a unified atlas has not been developed, there are several atlas projects underway that will link to other kinds of data – click on a particular label and launch a query.

Collection Interoperability – Databases and Repositories

There are a variety of technological and social barriers to the creation of large-scale interoperable systems. Successful implementation of these systems depends upon social agreements about (discipline-specific or specialized) standards for conducting (and thus describing) research activities, standards of validity and controlled vocabularies. Groups in most scientific communities are grappling with these issues and the processes of enacting their subsequent arrangements. Understanding how different disciplines communicate, interact with information resources and share data sets will help systems and metadata developers meet users' needs. However, it is likely that the following technological challenges may be less difficult to solve than many of the present social barriers.

Volume. The trend in the mid-90s was to try to build centralized databases for accumulation of research data. However, tools and methods have facilitated increasingly larger sets of data – geospatial data and brain imaging data can easily reach the terabyte level. Informaticists, computer scientists and domain experts have realized that centralized databases were not a viable approach to meet these scientific needs. Although computer storage devices continue to increase in capacity, they will not be able to hold everything. Further, disparate sets of data require different kinds of organization for management and querying. Distributed, federated systems will support long-term maintenance and allow faster and more flexible information retrieval (Jones, et al, 2001; Van Horn, et al, 2001).

Heterogeneity and Complexity. Biological data in its entirety represents concepts and structures from multiple domains that study the earth and its creatures along several dimensions. In addition, biodiversity researchers will need to use data from multiple new sources to develop future interpretations of species data. These disparate data types include data on land formations, climate changes, ecologies, animal and plant physiology and anatomy, biological systems, molecular and chemical biology, and behavior. Much of this data, collected by researchers and scientists in other fields such as climatology, was recorded for different purposes than to meet the analytic needs of biologists. Such variety presents a range of integration and re-use problems. By contrast in neuroscience, there is a singular purpose to the research – to make discoveries about the structure and function of the central nervous system, particularly the brain. However, the range of disciplines and research methods produce many forms of data, including numeric, textual, image and time-series. The resulting heterogeneity is not trivial, and this presents additional challenges to informatics researchers.

Item-Level Metadata. In the context of biological informatics, metadata is an interoperability issue. It is common to require metadata be submitted to databases along with research results, where it is tied to the validity and value of the scientific data itself. To preserve legacy data and increase the value of findings from current research, describing the experiments and data from that research is a necessity. The fMRI Data Center at Dartmouth, for example, has established four areas of metadata that depositors must include – protocol (about technique), human subject information, scan session information and experimental protocol (Van Horn, et al, 2001). This is a specific area where library and information scientists can contribute to the development of biological data collections, as our field is actively engaged in creating and testing solutions for metadata systems.

Access and Use

There is now consensus within biodiversity and neuroscience informatics fields that a distributed architecture with federation will best support the scale and complexity of the data. This approach will address the frequent concern about data control by keeping the location and control of these databases at their local institutions and countries.

Nomenclature. There are problems of nomenclature that will need to be solved. In biodiversity there are examples of multiple naming, where “new” discoveries of previously named species are given different names. In addition, millions of species remain unnamed. To name these new species we must sift through the tens of millions of named species to insure they were not previously named. Electronic systems support annotations and current rules for addressing this problem by making the different specimen types distinguishable. However, more will need to be done to support ongoing disambiguation, and there is a general lack of funds to support the basic taxonomy. Better global access to data sources like electronic flora and fauna will facilitate the information work that is required to verify either novelty or accurate identification of a specimen.

In neuroscience, there are at least two distinct problems with nomenclature. First, many regions of the brain do not have inherently or anatomically distinct boundaries. What Dr. Smith calls cell cluster "A," Dr. Jones includes in his identification of cell cluster "T." The second type of problem is related to synonyms; there may be multiple terms representing a single action or concept. Current IR systems are not very good at returning results across the range of terms. Calls for standardization have been met with resistance, as scientists are hesitant to give up the meaning that attaches to their particular terms. Library and information scientists specialize in these sorts of problems and could contribute here by helping to develop powerful thesauri and new retrieval tools.

Federating Data and Queries. Building collection-level metadata repositories will also help to support global research. These repositories will allow queries to be returned faster, as they will first find the collections from which the query might be answered. "Query-time federation" is being developed using at least two models – one that steers queries through a central metadata repository which then directs the query to (appropriate) databases for retrieval and the other using a mediating protocol such as Z39.50 where the client queries each of the repositories when the user executes a query.

Quality Assurance. Scientists and researchers want to have some trust in the reliability and specific application of their data and tools. Informatics projects are testing ways to support data quality, using variations of peer-review, annotation systems and cross-referencing, data weighting and others (Amari, et al, 2002; Chicurel, 2000). These systems have not yet seen systematic measurement, so it is too early to know which approaches will provide the best way to meet the needs of the various user groups. In biodiversity, the very act of digitizing museum specimens opens up the possibility of data validation. For example, scientists and others worldwide will be able to evaluate the validity of the name of the specimen using its location information and general knowledge of taxonomic revisions.

Conclusions

There are economical, political, social and technological challenges to bringing biological data together in a way that will permit new kinds of integrative analysis (see other articles in this issue). Technological and intellectual challenges include the lack of standards, developing and testing novel integration and retrieval techniques, and building thesauri and vocabulary systems. Information science can help solve the problems of managing and analyzing growing amounts of heterogeneous data. The movement in biological sciences toward global and systems level knowledge will require the sharing and integration of disparate data and tools that will support testing new kinds of questions. Finally, expanded study is needed to refine systems development, to illuminate "triggers" of collaboration and describe the implications for scientific practice and scholarly/scientific communication.

For Further Reading

Amari, S-I., Beltrame, F., Bjaalie, J.G., et al. (The OECD Neuroinformatics Working Group). (2002). Neuroinformatics: The integration of shared databases and tools towards integrative neuroscience. *Journal of Integrative Neuroscience*, 1 (2), 117-128.

Biomedical Informatics Research Network (BIRN). Available at http://www.nbirn.net/

Bowker, G. (2001a). Biodiversity, datadiversity. *Social Studies of Science, 30* (5), 643-683.

Bowker, G. (2001b). Mapping Biodiversity. *International Journal of Geographical Information Science, 14* (8), 739-754.

Chicurel, M. (2000). Databasing the brain. *Nature, 406,* (6798), 822-825.

Huerta, M. F., & Koslow, S. H. (1996). Neuroinformatics: Opportunities across disciplinary and national borders. *Neuroimage, 4* (3), S4-S6.

Jones, M. B., Berkley, C., Bojilva, J., & Schildhauer, M. (2001). Managing scientific metadata. I*EEE Internet Computing, 5* (5), 59-68.

Koslow, S. K. (2000). Should the neuroscience community make a paradigm shift to sharing primary data? *Nature Neuroscience, 3* (9), 863-865.

OECD Megascience Forum Working Group on Biological Informatics. (1999). *Final report of the OECD Megascience Forum Working Group on Biological Informatics.* Available at www.gbif.org/GBIF_org/facility/BIrepfin.pdf.

Toga, A. W. (2002). Neuroimaging databases: The good, the bad, and the ugly. *Nature Reviews Neuroscience, 3* (4), 302-309.

Van Horn, J. D., Grethe, J.S., Kostelec, P., Woodward, J. B., Aslam, J. A., Rus, D., Rockmore, D., & Gazzaniga, M. (2001). The Functional Magnetic Resonance Imaging Data Center (fMRIDC): The challenges and rewards of large-scale databasing of neuroimaging studies. *Philosophical Transactions of the Royal Society of London: Biological Sciences, 356* (1412), 1323-1339.

Special Section

The Role of Information Science in Gathering Biodiversity and Neuroscience Data

by Geoffrey A. Levin and Melissa H. Cragin

Geoffrey A. Levin is director of the Center for Biodiversity at the Illinois Natural History Survey, 607 E. Peabody Dr., Champaign, IL 61820-6970. He can be reached at 217-244-7481 or by e-mail: glevin@inhs.uiuc.edu

Melissa H. Cragin is a doctoral student in the Graduate School of Library and Information Science at the University of Illinois, Urbana-Champaign, 501 E. Daniel St., Champaign, IL 61820. She can be reached at 217-244-0866, or by e-mail: cragin@uiuc.edu.

As with the rest of science, biodiversity and neuroscience are becoming increasingly technological and data-rich. At the same time biodiversity studies retain many traditional tools and materials, while in neuroscience new tools are frequently developed. Information science can help both fields develop tools that exploit modern technologies to increase data-gathering efficiency, to improve quality control and, where necessary, to integrate historical and modern methods and materials. We will present overviews of biodiversity and neuroscience separately and conclude by summarizing how information science can help both fields address similar issues.

Biodiversity

A valuable perspective on biodiversity studies can be gained from considering the Lewis and Clark Expedition, whose bicentennial the United States will celebrate in 2004. Like explorers before and after them, they and their Corps of Discovery were charged not only with exploring and mapping the land they crossed, but also with documenting the plants, animals, minerals and indigenous peoples they encountered (Ambrose, 1996). They collected numerous specimens and artifacts that they removed from the region of origin and sent to institutions and individuals in the "developed" world, in this case the northeastern United States. Most of the plant specimens, for example, are at the Academy of Natural Sciences in Philadelphia (www.acnatsci.org/museum/lewisclark/l&c_herbarium.html). There, experts studied the specimens and described new species, often over a span of many years. Unlike most other 19th century explorers Lewis and Clark kept extensive notes that still survive, so it is possible to know fairly precisely where the specimens came from, their ecological setting and other relevant information. More typical are notes like those of mid-19th century explorer Charles Wright, many of whose specimens, distributed to museums in the eastern United States and Europe, are accompanied only by notes giving the month and year he collected and the location as "western Texas."

Collecting practices like these continued well into the 20th century, persisting longest in tropical areas. In well-explored areas, collectors were able to give precise locations, but in tropical areas they may only have known the river they were along and how many hours of paddling it took to get there from a distant village. Ecological information with the specimens tended to be idiosyncratic, with some scientists providing good descriptions of vegetation, soil and other relevant conditions, whereas others provided little or no data. Beginning in the 19th century, the most significant change was the elimination of the amateur collector. Instead, predominant practice throughout the 20th century has been for professional scientists to serve both as field collectors and museum experts. Museums worldwide now hold an estimated two to three billion biodiversity specimens, about 75% in the industrialized countries, and the number continues to grow (www.gbif.org/GBIF_org/facility/BIrepfin.pdf).

There are standard practices for treating and preserving new specimens. Plants generally are flattened, dried and glued to reinforcing archival paper. Vertebrate animals may be preserved whole in alcohol or represented only by their skins, bones or shells. Invertebrates like insects are pinned, preserved in alcohol or mounted on microscope slides. In all cases specimens are accompanied by a label or tag with information about the collector, collection location and date and sometimes by additional information about the habitat or characteristics of the organism. Although the analogy is not perfect, scientists often liken scientific collections to libraries, although books can be reproduced whereas each specimen is unique. Techniques for managing specimen data and book cataloging information are similar; historically specimen data often were maintained in paper catalogues, whereas these data now are being placed in databases, often with Web access (see other articles in this issue).

In recent years the nature of biodiversity specimens has changed, particularly as a result of the application of molecular biological techniques to biodiversity research. Standard specimens now often are accompanied by specimens preserved in ways that protect the organism's DNA, such as freezing, special chemical treatment or rapid desiccation. The DNA itself may be isolated and preserved (usually frozen), and portions may be sequenced and the data deposited in a central database (e.g., GenBank, www.ncbi.nlm.nih.gov/Genbank/index.html). Fungi, protozoans and bacteria frequently are maintained as living cultures, and other living collections like zoos and botanical gardens are sometimes managing their collections as consisting of biodiversity specimens. Frequently a single individual may be represented by several specimens, e.g., a typical museum specimen, frozen tissue and several sequenced DNA regions, necessitating record keeping that retains the connection among the specimens and the associated data, often all in different locations.

Even more significant are changes in the types of data being gathered. In the past, the focus was principally on gathering representative specimens, whereas now the focus often is on the whole ecosystem. Inventories are receiving renewed attention, but they are increasingly quantitative and are often driven by societal needs, such as ecological classification for land use decision-making (including rapid assessments for conservation) and ecosystem health monitoring. Because taking specimens is costly in terms of time, material and storage space and generally requires destructive sampling, there has been an increase in the use of observational data only. Without specimens that can be used to verify identifications, obvious quality assurance issues arise.

There has also been a resurgence in data gathering by trained amateur "citizen scientists" as the demand for inventories outstrips the number of professional scientists available (see, for example, the Illinois Department of Natural Resources EcoWatch Program; dnr.state.il.us/orep/ecowatch/index.htm). A particularly ambitious example of enlisting amateurs is eBird (www.ebird.org), which aims to utilize the popularity of bird watching to compile an extensive bird census database for North America. Quality assurance needs that arise with such data can be addressed in various ways. Developing methods for delivering identification tools to the field, possibly augmented by real-time connections with professional scientists in the laboratory, can improve quality assurance. Integration of quality control tools directly into the data gathering process offers still further improvement. For example, eBird responds with specific identification queries when rare, difficult-to-identify or out-of-range species are entered into the database. Field data entry using handheld computers linked through satellites to remote databases could further increase data gathering efficiency for both amateurs and professionals and could even be used to generate labels for specimens as they are collected.

In addition to using observational data on organisms, biodiversity studies are increasingly integrating other forms of data. Locations are now being determined using the Global Positioning System (GPS), providing high precision and accuracy. This facilitates integration with remote sensing, mapping and other spatial data through Geographic Information Systems (GIS). Increasingly, quantitative environmental data are also being collected. The United States Longterm Ecological Research (LTER) Network (lternet.edu), which includes 24 sites in different ecosystems, has been collecting such data for about 20 years. These efforts would expand tremendously with the proposed U.S. National Science Foundation National Ecological Observatory Network (NEON). It is estimated that the NEON, when completed, would generate more than 20 million observations daily (www.sdsc.edu/NEON). The demand for efficient data gathering and management will necessitate the participation of information scientists in the planning and execution of the NEON and similar efforts worldwide.

The history of biodiversity data gathering thus shows a significant shift in practice. What was once the work of individuals or small groups working largely in isolation has grown to involve multidisciplinary collaborative projects involving large numbers of people and volumes of data. Field sites are becoming increasingly connected to each other and to scientific institutions worldwide through modern telecommunication links.

Neuroscience

Such significant shifts in scientific practice are equally visible in neuroscience. The application of networking and computational techniques is leading to new forms of collaboration and amalgamation of data. A hundred years ago, data were collected from medical patients and published in the form of case studies. Other data were collected using cadavers' brains or through in vitro animal experiments in which researchers would introduce brain lesions and record the results. (Star, 1989. Regions of the Mind: Brain research and the quest for scientific certainty. Stanford, CA: Stanford University Press) Data content included notations on symptoms, treatment information and post-treatment functionality, while data consisted of text and graphics.

Data gathered on the brain during the last several decades has shown significant change, mainly because of changing tools and techniques. Microscopes and electrophysiology systems were improved and refined. The former allow collection of greater amounts of anatomical data for smaller and increasingly specific parts of the brain. The latter allow electrophysiologists to record multiple trials of electronic "firings" of neuron cells in response to some stimuli, rendering numeric time-series data. Together with scientific developments, these technological innovations have reinvigorated neuroscience, and this is evident in the vast amount of published research. In 1997, the *Journal of Neuroscience* literally doubled its production rate, expanding to two issues per month; in 2001, ISI included 198 titles in its neurosciences journal group; and in 2002-2003, at least two new journals specific to neuroinformatics began publication.

This indication of growth has not been limited to publishing, but is apparent in frequent modification of methods and the increase in the amount of data resulting from the ability to observe and analyze smaller and smaller parts of the nervous system. Work in genetics moved the field toward molecular biology, and electrophysiology and scanning techniques have improved several fold. Neuroscience is a large and diverse field, and there is great variability in research incorporating anatomy, neurophysiology, molecular biology, chemistry and computational modeling. Neuroscience data are gathered in several forms, including numeric, textual, image, graphic and time series. Behavioral and cognitive constructs studied include attention, memory, learning, perception, emotion and language. Research on brain function produces data collected across a number of dimensions, including

- molecular to synaptic, and cellular to systemic;
- single organisms to populations;
- normal through various states of disease;
- birth through old age; and
- across species.

It is understandable that neuroscience investigations often produce large sets of data, but it is easy to overlook the fact that only part of such results may ever be analyzed and reported in a paper. Historically, this data has been collected and stored locally, and there was no tradition of depositing data for future or external use. There is little value in legacy data, which results in years of neuroscience data being disregarded or even lost over time. It should be noted that there is also undeposited data in the biodiversity field; there are many collections "orphaned" as scientists retire. The need for strategies to archive and preserve data is an opportunity for LIS practitioners to engage with scientists to find solutions. It will be particularly important to develop guidelines about what data is important to save, as there is no consensus about this approach (Chicurel, 2000). Research results and new claims continue to be published in traditional ways, and there is growing concern about the loss of biological data going forward (Valencia, 2002).

The emergence of *neuroinformatics* in the 1990s has created new possibilities for discovery, but with this there has also been a shift in the public focus, funding and expectations of scientists, and there has been concomitant increase and modification of data collection. The Human Brain Project (HBP) (www.nimh.nih.gov/neuroinformatics/index.cfm) in the United States was started in 1993, following the report of the Institute of Medicine on the need for a coherent agenda for brain research (Pachura and Martin, 1991). From the outset, one of the goals of HBP has been to combine informatics research with neuroscience research. As noted above, biologists studying biodiversity are recognizing the benefits of bringing different and disparate data together; this is beginning in neuroscience. Today HBP supports research projects to federate data queries, refine viewing of cell structures and develop imaging databases that will support meta-analysis and other future use. All of these projects work to bring raw data and supporting materials (such as annotations) together. Other studies have focused on developing systems that store and link textual and image data or computational modeling with neuronal data. (See for example, the fMRI Data Center at Dartmouth, www.fmridc.org; the SenseLab at Yale, http://senselab.med.yale.edu/senselab/)

For Further Reading

Ambrose, S. (1996). *Undaunted courage: Meriwether Lewis, Thomas Jefferson, and the opening of the American West.* Simon & Schuster.

Chicurel, M. (2000). Databasing the brain. *Nature, 406* (6798), 822-825.

ebird. Available at www.ebird.com.

fMRI Data Center. Available at www.fmridc.org.

Human Brain Project website at www.nimh.nih.gov/neuroinformatics/index.cfm.

Lewis & Clark Herbarium website at www.acnatsci.org/museum/lewisclark/l&c_herbarium.html.

Pachura, C.M. & Martin, J.B. (Eds.). (1991). *Mapping the brain and its functions: Integrating enabling technologies into neuroscience research.* Washington, D.C.: National Academy Press.

SenseLab at http://senselab.med.yale.edu/senselab/.

Star, S. L. (1989). *Regions of the mind: Brain research and the quest for scientific certainty.* Stanford, CA: Stanford University Press.

OECD Megascience Forum Working Group on Biological Informatics. (1999). *Final report of the OECD Megascience Forum Working Group on Biological Informatics* at www.gbif.org/GBIF_org/facility/BIrepfin.pdf.

Valencia, A. (2002). Search and retrieve. *EMBO Reports, 3* (5), 396 – 400.

Working group reports on the *National Ecological Observatory Network* at www.sdsc.edu/NEON.

Shared Problems

The idea of sharing or making public one's data is central to biodiversity informatics and neuroinformatics, and this conflicts with traditional scientific practices. Scientists have voiced a variety of concerns about sharing data; two are widely recognized. Scientists and human subjects (patients or tissue donors), corporations and political entities all need to be assured of the reliability and security of data being deposited in databases and repositories. Neuroinformatics projects that serve as databanks or repositories implement a variety of procedures to improve data quality. The most common tactic is to only accept data from studies that have gone through a peer-review process of a submission to a journal. Another approach is required submission of a set of augmentative information about the experiment or data so that others can judge its value with respect to later use. This information might include, for example, settings on an imaging scanner and the paradigm used for the experimental design. There is some debate among biologists and other researchers about what parts of this material constitute metadata. However, gleaning and managing the significant information related to shared data is important and necessary for both biodiversity and neuroscience.

Although biodiversity studies and neuroscience differ significantly in scope and focus, both will depend on information science to provide needed expertise if they are to effectively address scientific and societal needs. People doing many different types of work utilize biological specimens. This variety is not so much the case in neuroscience. For obvious reasons amateur "citizen scientists" are not involved in neuroscience. Yet even though the heterogeneity of biodiversity data is qualitatively different from neuroscience data, there are many ways that these fields would similarly benefit from the traditional knowledge of library and information science. Information science researchers and practitioners can bring expertise in data visualization and retrieval techniques, records management, quality assurance and usability.

The Global Biodiversity Information Facility

by Meredith A. Lane

Meredith A. Lane is a public relations and communications officer with the Global Biodiversity Information Facility (GBIF) Secretariat. She can be reached at Universitetsparken 15, 2100 København Ø, Denmark; mobile: +45 28 75 14 84; phone: +45 35 32 14 84; fax: +45 35 32 14 80; e-mail: mlane@ gbif.org.

"B*iodiversity* or *biological diversity* means the variability among living organisms from all sources including, *inter alia*, terrestrial, marine and other aquatic ecosystems and the ecological complexes of which they are part; this includes diversity within species, between species and of ecosystems" (Convention on Biological Diversity, Art. 2, para. 1).

The Global Biodiversity Information Facility (GBIF) (www.gbif.org) mission *is to make the world's biodiversity data freely and universally available via the Internet.* GBIF was established and is being maintained by an ever-growing consortium of forward-thinking countries, economies and international organizations. Its charge is to carry out specific tasks that are essential to a worldwide infrastructure that can overcome current barriers to the universal availability of species-level biodiversity information.

The GBIF Vision

Science. Significant parts of the tripartite (gene, species, ecosystem) biodiversity information resource are already online, such as the DNA data served by GenBank, EMBL and DDBJ and other sequence (RNA, protein, etc.) data served by various sources such as RNAbase, SCOR and ExPASy. These community research resources have already contributed substantially to medical, pharmaceutical and agricultural industries and through these to society, which has footed the bills for the establishment and maintenance of the digital information resources.

Ecological, ecosystem and planet-wide data are provided online by, for example, MABnet and LTERnet, the U.S. National Aeronautics and Space Administration (NASA) through its Mission to Planet Earth (MPE) and other national and bi- or trilateral consortia. These online ecological resources are just beginning to be thoroughly analyzed and the results synthesized into larger understandings of the functioning of the natural systems of our planet. Nonetheless, they hold great promise for predictive modeling of global climate change and other large-scale ecological phenomena.

Though there are areas of the world that still do not have either full access to the Internet or the on-site infrastructure to fully utilize data that might be received (one hopes that this situation will be rectified in the near future), the molecular and ecological data discussed above are available to anyone and everyone with an Internet connection. One of the reasons that this is so is that the available data have all been collected in the past few decades, during the age of computers.

The piece, in fact the anchor stone, of biodiversity information that to-date has been missing from digital availability is data about individuals, populations and species of whole organisms. These data are not online because most pre-date the computer age. They have been collected over nearly three centuries and recorded in the only medium available: paper and ink. These data are on the labels of natural history specimens, in libraries and in handwritten notebooks or typewritten card files. Species-occurrence data are essential to many kinds of analyses, and one of GBIF's four major areas of

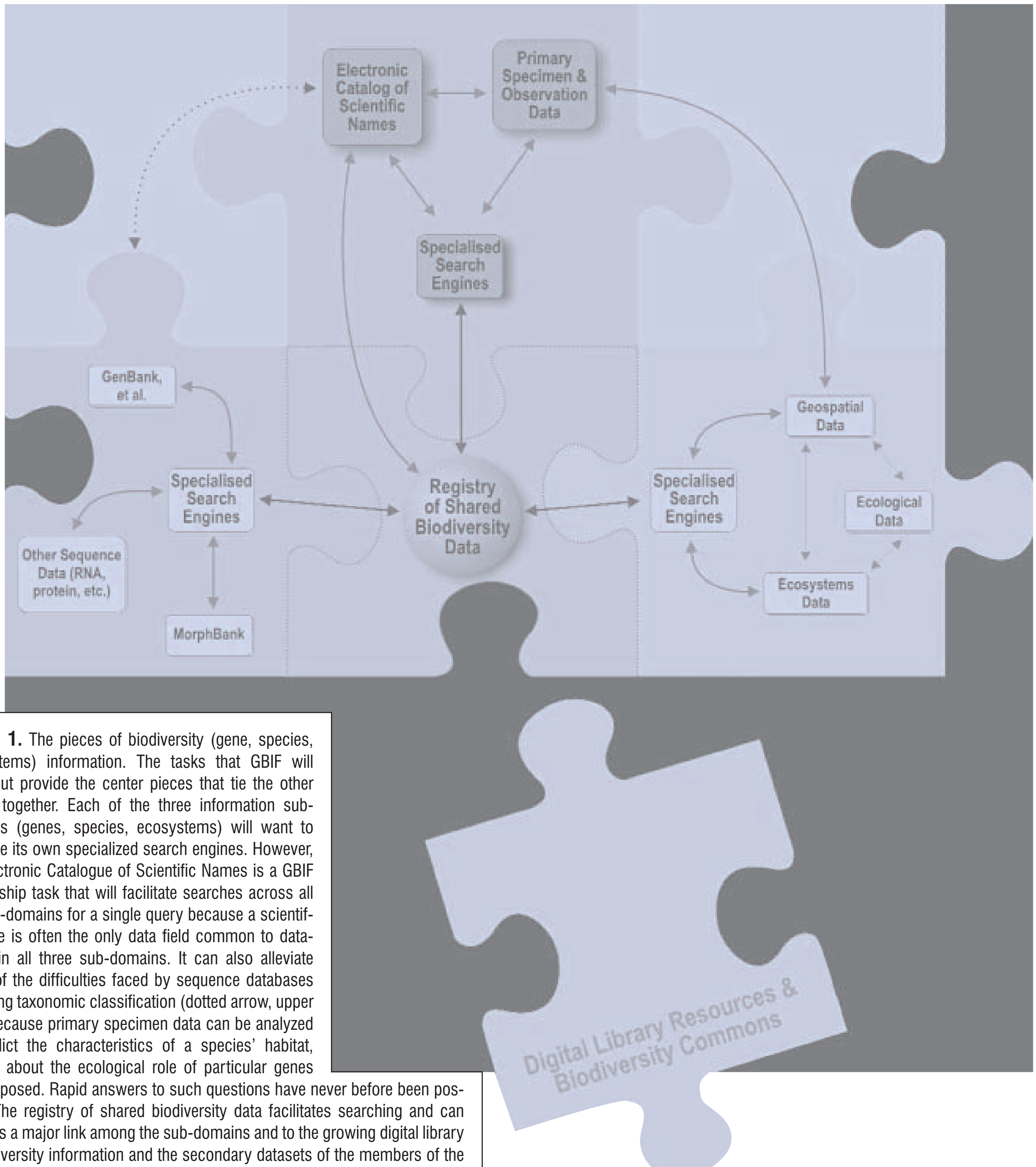


Figure 1. The pieces of biodiversity (gene, species, ecosystems) information. The tasks that GBIF will carry out provide the center pieces that tie the other pieces together. Each of the three information sub-domains (genes, species, ecosystems) will want to generate its own specialized search engines. However, the Electronic Catalogue of Scientific Names is a GBIF partnership task that will facilitate searches across all the sub-domains for a single query because a scientific name is often the only data field common to databases in all three sub-domains. It can also alleviate some of the difficulties faced by sequence databases regarding taxonomic classification (dotted arrow, upper left). Because primary specimen data can be analyzed to predict the characteristics of a species' habitat, queries about the ecological role of particular genes can be posed. Rapid answers to such questions have never before been possible. The registry of shared biodiversity data facilitates searching and can serve as a major link among the sub-domains and to the growing digital library of biodiversity information and the secondary datasets of the members of the Biodiversity Commons. The first-generation GBIF information architecture will be online by the end of 2003; the growth of the electronic catalogue is underway, but is expected to take approximately 10 years to achieve 90% of all names. Digitization of natural history specimens has begun, but needs substantially more investment to speed the work. (Source: GBIF)

work is aimed at making these data digital and therefore more useful and available.

All of GBIF's tasks are aimed at making digitally available the bridging pieces (data about individuals, populations and species) that tie the whole biodiversity information pattern together (see Figure 1). Without the Taxonomic Name Service (a function that uses both the Registry of Shared Biodiversity Information and the Electronic Catalogue of Scientific Names) that will be provided by GBIF, there is no way to obtain a seamless response to a query that requires a call upon both ecological and molecular data, such as *"What other organisms live in the same kinds of habitats as this one from which I have extracted a gene that enables it to tolerate high levels of lead acetate, and are able to do so because of either homologous or analogous gene function?"*

Society and Sustainability. The beauty of digital data is that they can be used over and over again and in many ways. The same information can be put to many purposes by the whole variety of users in the world. Investments made in digitizing scientific data and bringing them online via well-planned information architecture are paid back many times over. Not least of these repayments is that these data are now usable by people other than scientists.

The ultimate source of global and national wealth is natural resources. Biodiversity, the living portion of natural resources, provides an ever-increasing portion of that wealth. Biodiversity has provided the basis of human survival (clean water and air, food, fuel and fiber), not to mention prosperity, since *Homo sapiens* first set foot on Earth. If we humans are to continue to prosper and to leave future generations a healthy place in which to live, we must learn to use living resources in a sustainable way and make it possible for all peoples to share in the benefits of the sustainable use of biodiversity.

If the economic and survival benefits provided by biodiversity are to be equitably shared globally, as required by the Convention on Biological Diversity, a number of factors must be overcome. Primary among these is the need for access to scientific data and information about biodiversity to be as easy and complete in Mongolia or Madras as it is in Madrid or Munich. If scientific advances regarding biodiversity are to be made by any and all of the talents around the world that can contribute to them, access to the data and results that have already been generated must be as easily and fully available, wherever the researcher.

The GBIF Strategy

The GBIF biodiversity informatics infrastructure has two equally important components:

1) Computational (standards, interoperability and search engines)
2) Content (particularly primary scientific data that are currently "imprisoned" by paper and ink or other, non-digital media)

GBIF is also aware of the digital divide faced by many of the most biodiverse countries and, therefore, undertakes capacity-building activities to help overcome these challenges.

At present GBIF is focusing on making primary, species-level biodiversity data available via the Internet. Primary data are those derived from the direct observation of nature, such as the labels on natural history specimens or culture tubes. Eventually, GBIF will include in its purview secondary data which have been derived through some manipulation of primary data, such as an analysis of pattern or process. This task will likely be done in close collaboration with digital library efforts, as well as the Biodiversity Commons. However, the immediate need is to provide, via the Internet, the label data from more than two billion natural history specimens and uncounted culture collections.

GBIF is an international organization in its own right. Importantly, GBIF is open to participation by *all* countries, economic entities and organizations that can benefit from the open sharing of biodiversity information on a global scale.

Unlike other megascience facilities that are built of *bricks* and *mortar*, GBIF operates as a virtual facility. The bricks of this facility are the databases, other information resources and informatics tools made available by GBIF participants. The mortar that holds the bricks together is the informatics infrastructure (software tools and the Internet). It is staffed by a small secretariat (14 positions) that works internationally to coordinate national, regional and local biodiversity informatics efforts and bring focus to the activities of the organization as it develops GBIF.

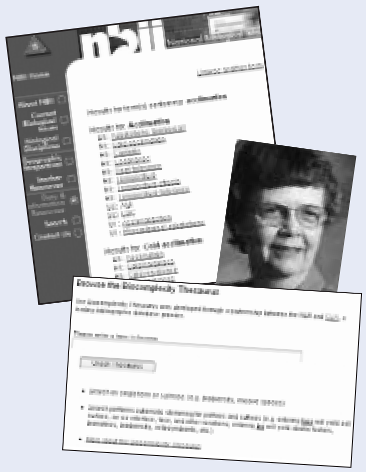
GBIF is distributed and encourages cooperation and coherence; it is global in scale, though implemented regionally, nationally and locally. GBIF, through relationships built by its secretariat staff, works closely with many producers and providers of biodiversity information as well as with its users. In some ways GBIF has the characteristics of a large, distributed public-domain database with a number of interlinked and interoperable modules such as data stores, software and networking tools, search engines and analytical algorithms that enable users to navigate and use the data. It differs from such a unitary database, however, in that it is more comprehensive in content and much more complex in its interconnections.

In summary, the GBIF strategy includes:

- Focus on its mission and specific goals, with intermediate milestones identified in each year's work program;
- Outreach to developing countries;
- Inclusiveness in the manner in which it seeks advice;
- Openness in data sharing and software developments;
- Cost-effectiveness in its partnerships with like-minded organizations; and
- Fund-raising efforts to enhance its product and speed up its activities.

Biocomplexity Thesaurus Launched

Reprinted from U.S. Geological Survey. (2003). NBII Access, newsletter of the National Biological Information Infrastructure, 6 (3).



On May 20, 2003, the National Biological Information Infrastructure (NBII) and CSA (Cambridge Scientific Abstracts) announced the launch of the Biocomplexity Thesaurus, a major new resource for the bioinformatics community. The thesaurus is now online and available for use by the NBII nodes, their partners and the public-at-large. The Biocomplexity Thesaurus can be accessed at http://thesaurus.nbii.gov.

The Biocomplexity Thesaurus will be integrated into NBII products and services to facilitate more relevant retrieval of and intellectual access to these resources by NBII users. It will be the required thesaurus for all keyword and subject metadata created by NBII nodes. The thesaurus will be used for the cataloguing of Web resources, the creation of HTML metadata for Web pages and the creation of new Federal Geographic Data Committee (FGDC)-compliant metadata records for the NBII Metadata Clearinghouse. Accordingly, researchers, scientists, librarians and the general public will be able to use the thesaurus as an aid in creating strategies for searching NBII databases.

This thesaurus represents a major step forward for the biological sciences and the NBII network. It is one of the most comprehensive and freely accessible biological thesauri available online.

The Biocomplexity Thesaurus is a living resource that will be updated quarterly. To this end, the NBII Thesaurus Working Group (TWG) has been established to review recommended additions and modifications to the thesaurus. The goal is to have representation from every node on the TWG. Two NBII regional nodes, Pacific Basin Information Node and Southern Appalachian Information Node; CSA, the thesaurus developer; and the NBII Program Office already have representation on this working group. Representatives from the thematic nodes and the other regional nodes are encouraged to participate as well.

Additionally, a Biocomplexity Thesaurus community has been set up in the NBII portal. All nodes may submit requests for term additions or modifications through a gadget located in this online community. NBII node partners may also submit comments and requests through this gadget. Contributors need not be members of the TWG to access this Community.

Jessica L. Milstead is the primary architect of the Biocomplexity Thesaurus. She is one of the foremost experts in the world on thesaurus development for the biological and natural sciences. She holds a doctorate in library science and is the founder of the index and thesaurus development company, JELEM (www.jelem.com). Dr. Milstead has extensive experience in both industry and academe. Under contract to CSA, she brought the massive Biocomplexity Thesaurus project to fruition in just 18 months instead of the 24 months originally allocated.

To create the Biocomplexity Thesaurus, Milstead merged, vetted and reconciled the terminology in five large existing thesauri plus one smaller specialty thesaurus collectively covering the biological, environmental, aquatic, ecotourism and sociological sciences. These thesauri include

- CERES/NBII Thesaurus (California Environmental Resources Evaluation System)
- CSA Life Sciences Thesaurus
- CSA Pollution Thesaurus
- CSA Aquatic Sciences and Fisheries Thesaurus
- CSA Sociological Thesaurus
- CSA Ecotourism Sciences Thesaurus.

Milstead used the high-powered MultiTes 8.0 thesaurus development software package to create the Biocomplexity Thesaurus. Among many other features, MultiTes provides for an unlimited number of hierarchies per thesaurus, an unlimited number of relationships for each individual term, the validation of conflicting relationships and the automatic generation of reciprocal relationships.

The online Biocomplexity Thesaurus is very user-friendly. Tips are offered for easily navigating and searching the thesaurus. Explanations are provided for best understanding search results, such as concepts related hierarchically or associatively to the search entry term.

The NBII (www.nbii.gov) is a broad, collaborative program to provide increased access to data and information on the nation's biological resources. The NBII links diverse, high quality biological databases, information products and analytical tools maintained by NBII partners and other contributors in government agencies, academic institutions, non-government organizations and private industry.

CSA (www.csa.com/csa/about/Biocomplexity.shtml) is a leading producer of bibliographic citation databases and Web resources databases. CSA recently extended its government/private industry partnership with the NBII from 20 to 56 months. CSA's original designation as the "Biocomplexity Information Node" has now been broadened to "NBII Infrastructure and Knowledge Integration Node."

Citation Analysis and Research Assessment in the United Kingdom

by Julian Warner

Julian Warner serves as international liaison to the ASIST Board of Directors. He is affiliated with the School of Management and Economics at The Queen's University of Belfast, Northern Ireland. He can be reached by e-mail at j.warner@qub.ac.uk

Evaluation of publicly funded university research in the United Kingdom has been conducted through a series of Research Assessment Exercises (RAEs) in 1992, 1996 and 2001. The results of the RAEs have been used as a factor in the calculation of public funds for distribution to universities for research. The primary mechanism used for the research evaluation has been direct peer review of submissions and publications from university departments. The conduct of the RAEs, the public debate generated and a recent central government report of future research evaluation has generated issues relevant to other jurisdictions and to the interest in research evaluation within information science. In particular, a crucial transformation in the value of citation analysis in research evaluation can be detected.

Two antithetical viewpoints on the appropriate relation of citation analysis to the RAEs have coexisted within information science, developing since the mid-1990s. The dominant approach established strong correlations between rankings of entities for research assessment derived from citation analyses and from RAE grades. Replacement of the process of peer review embodied in RAE procedures by an ordering (and conversion of the ordering into grades) determined by the results of citation analyses was then advocated on the grounds of broad equivalence of results and lower relative costs.

An alternative perspective (articulated by the current author) acknowledged the correlations established but argued that citation analysis may be used to inform, rather than to determine, judgment and does not even have to be used for this purpose. Advocacy of the use of citation analysis to determine grades has been strongest within information science as reflected in the number of authors and publications advocating it, the status of journals for publication and in numbers of citations received – although citations have been made predominantly from within information science rather than from other disciplinary literatures. Dialogue between the two viewpoints has been apparent but has not resulted in an accepted synthesis.

The dialogue could be regarded as occurring between incommensurable paradigms or at least as not being fully resolvable – from a Kuhnian perspective – by the participants from either paradigm. Resolution by an external observer, acknowledged to have cognitive authority for this purpose and concerned not only with the internal or logical coherence of each approach but with their real world effects and value, would then be welcome. An analogy could be made with court judgments between disputants and further with the process of referral to higher courts.

The recent central government report, *Review of Research Assessment*, was conducted by privileged observers credited with cognitive authority and concerned with real world implementation. The report does not directly engage with the literature of information science (and is free from a scholarly apparatus of citations and bibliography), but was informed by submissions, the understandings of participants in workshops and published literature on the RAEs.

Citation Analysis and Research Assessment

The congruence between the position developed by the critique of citation analysis and the recommendations of the report for future research assessment is striking. The critique had concluded

> *citation analysis can…be employed as one element used to inform judgment of research quality, with judgment underdetermined by any single element.*

Similarly, the first recommendation of the report insists:

> *Recommendation 1. Any system of research assessment designed to identify the best research must be based upon the judgment of experts, who may, if they choose, employ performance indicators to inform their judgment.*

Performance indicators should be understood to include the results of citation analyses. There are indications of movement over time in the understandings of some of the members of the steering group of the review, from a determining to an informing use of performance indicators.

Some of us believed, at the outset of the process, that there might be some scope for assessing research on the basis of performance indicators – we are now convinced that the only system that will enjoy both the confidence and the consent of the academic community is one based ultimately upon expert review.

The preference for peer review over the determining use of performance indicators is based principally on its ability to resist behavioral distortions and does acknowledge the contrasting levels of labor and costs involved.

The risk of the unproductive distortion of research behavior to conform with evaluative measures forms a persistent theme of the report and is emphasized in the Preface:

More important, I urge the funding councils to remember that all evaluation mechanisms distort the processes they purport to evaluate.

Only peer review can detect and resist unwanted behavioral distortions:

We are also convinced that only a system based ultimately upon expert judgment is sufficiently resistant to unintended behavioral consequences to prevent distorting the very nature of research activity.

Bibliometric measures are understood to be included in the performance indicators, which might promote undesirable results. The possibility of distortion of citation practices by the use of citation analysis in evaluation had been anticipated by the critique and by other commentaries.

The contrasting levels of direct human labor and the costs of that labor involved in peer review and in the determining use of performance indicators are acknowledged. The burden of peer review is accepted by the report. It is also recognized that the promise of reduced labor offered by citation analysis might be betrayed by the magnitude of the task of editing data into a form acceptable for comparisons between entities for assessment in a large country, such as the United Kingdom. The argument from relative costs for the use of citation analysis to determine judgment has then been considered and rejected.

The report's own and proposed use of citation analysis and other performance indicators is consistent with its informative role. In relation to current considerations, the increase in citations received by United Kingdom papers since the introduction of the RAEs is used to support the review's judgment of the improvement in United Kingdom research. For future use, peer reviewers are to be informed by performance indicators but not compelled to reflect them in grades awarded (for the Research Quality Analysis proposed to replace the RAEs). Performance indicators would only form the basis of assessment for those entities with quantities of research below the level commensurate with the costs of peer review (the Research Capacity Analysis).

Conclusion

A dispute within information science has been decisively resolved by privileged external observers, in favor of the view that citation analysis may be, and does not have to be, used to inform expert judgment. The congruence in conclusions between the report and the alternative perspective within information science is matched by similarities in reasoning, particularly the recognition of the risk of distortion of behavior. The report's preference for peer review is partly a product of changes in the reviewers' understandings, through their consideration of relevant evidence. The resolution of the dispute must be acknowledged by future work in citation analysis, if it is to retain its relevance to wider public discourse.

What's New?

Selected Abstracts from *JASIST*

Editor's note: We invite *JASIST* authors to submit structured abstracts of their articles for possible inclusion in the *Bulletin*, particularly those that might be of interest to practitioners. ASIST would welcome reader feedback on the usefulness of this (or any other) *Bulletin* feature (bulletin@asis.org).

FROM *JASIST* V. 54 (9)

Bergman, O., Beyth-Marom, R., Nachmias R. (2003). The user-subjective approach to personal information management systems, pp. 872-878.

Study and Results: Personal computers often serve as personal information management (PIM) systems: they allow people to collect items of information and store them outside their cognitive system. Research has shown that users of these systems find it hard to remember where they placed their personal information and thus have difficulties in retrieving it whenever necessary. In the article we propose a user-subjective approach to PIM system design that advocates that PIM systems should relate to the subjective attributes that users give to the data stored in them. This will hopefully facilitate system use: help users find information items again, recall them when needed and use them effectively in subsequent interactions with them.

Whats New? Three generic principles for system design are suggested and discussed: (a) The *subjective classification principle* stating that all information items related to the same subjective topic should be classified together regardless of their technological format; (b) the *subjective importance principle* proposing that the subjective importance of information should determine its degree of visual salience and accessibility; and (c) the *subjective context principle* suggesting that information should be retrieved and viewed by the user in the same context in which it was previously used. We claim that these principles are only sporadically implemented in operating systems currently available on personal computers and demonstrate alternatives for interface design.

Limitations: Empirical data are presently collected to support our claims, but are not yet reported in this article.

FROM *JASIST* V. 54 (10)

Borlund, P. (2003). The concept of relevance in IR, pp. 913-925.

Study and Results: The concept of relevance is introduced as viewed and applied in the context of information retrieval (IR) evaluation and presented in accordance to the multidimensional and dynamic nature of the concept. The literature on relevance reveals how the concept is many-faceted. The multidimensionality of relevance does not refer only to the various relevance criteria users may apply in the process of judging relevance of retrieved information objects. The multidimensional aspect of relevance covers also classes, types and degrees of relevance. Further, special attention is paid to situational relevance, which is discussed with reference to its potential dynamic nature and as a requirement for interactive IR evaluation.

Whats New? It is often said that no consensus exists on the relevance concept. However, the outlining of the multidimensional and dynamic nature of relevance provides a framework, which demonstrates that a consistent and compatible understanding of the relevance concept has been reached at an overall level.

Limitations: The focus of the paper is limited to the application of relevance within the research area of IR.

Hara, N., Solomon, P., Kim, S., & Sonnenwald, D. H. (2003). An emerging view of scientific collaboration: Scientists' perspectives on collaboration and factors that impact collaboration, 952-965.

Study and Results: This paper describes collaboration among a group of scientists and considers how their experiences are socially shaped. The scientists were members of a newly formed distributed, multi-disciplinary academic research center. To investigate challenges that emerge in establishing scientific collaboration, data were collected about members' previous and current collaborative experiences, perceptions regarding collaboration and work practices during the center's first year of operation. The data for the study includes interviews with members of the center, observations of videoconferences and meetings, and a center-wide sociometric survey.

Whats New? Data analysis has led to the development of a framework that identifies forms of collaboration that emerged among scientists (e.g., complementary and integrative collaboration) and associated factors, which influenced collaboration including personal compatibility, work connections, incentives and infrastructure. Practitioners can learn about possible issues to attend to in their own developing collaboration situations, especially collaboration in dispersed locations. These results may inform the specification of social and organizational practices, which are needed to establish collaboration in distributed, multi-disciplinary research centers.

Limitations: This study was of an emerging scientific collaboration. The purpose was to understand facilitators and barriers to collaboration as seen by participants in this collaboration. The findings may not apply to other settings. The resulting framework is offered to enable consideration of applicability in other settings as well as to inform other research on factors impacting on collaboration.