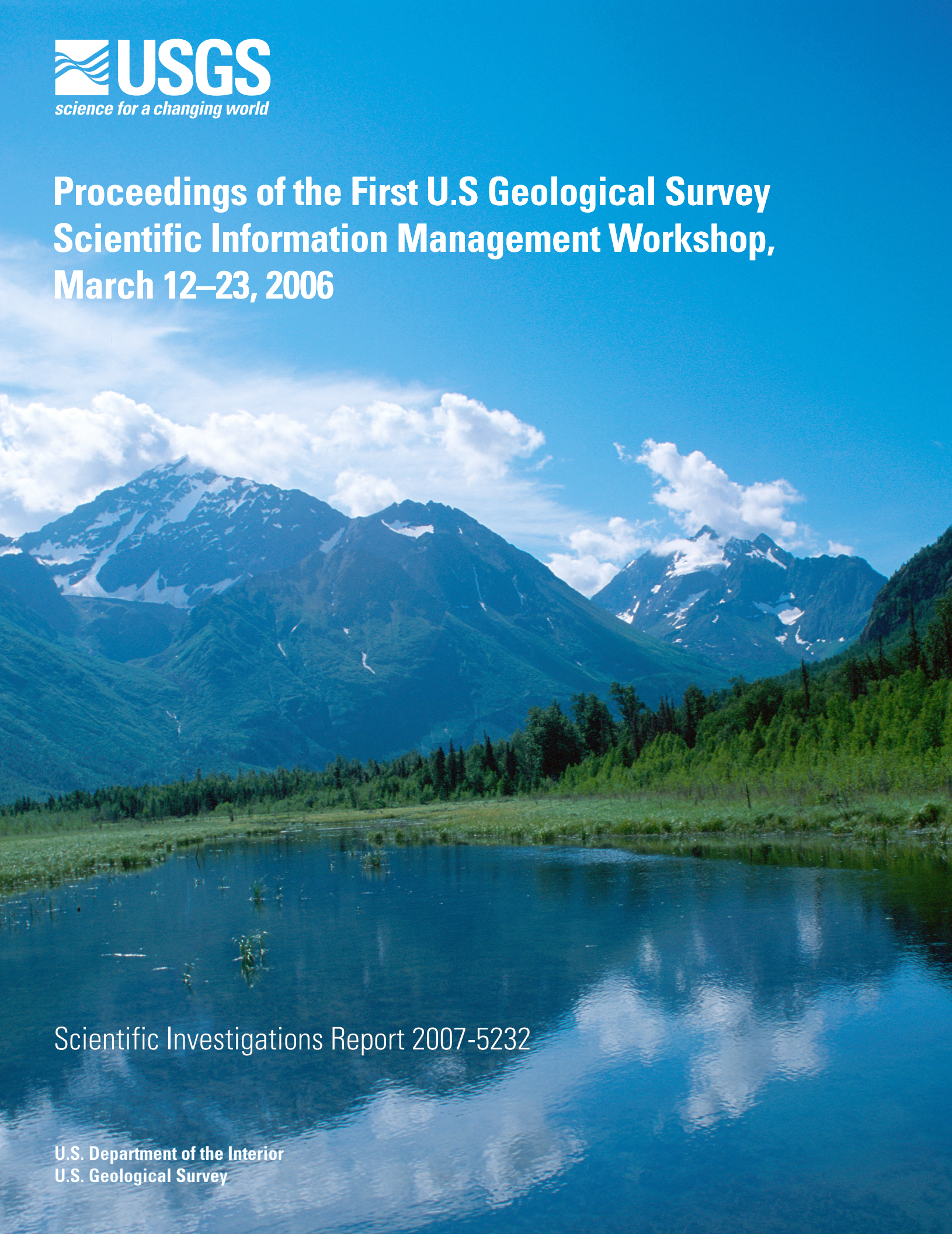
USGS
science for a changing world

Proceedings of the First U.S Geological Survey Scientific Information Management Workshop, March 12–23, 2006



Scientific Investigations Report 2007-5232

U.S. Department of the Interior
U.S. Geological Survey

USGS
science for a changing world

Proceedings of the First U.S. Geological Survey Scientific Information Management Workshop, March 21–23, 2006

Compiled by Heather S. Henkel

Sponsored by the Coastal and Marine Geology Program, the Enterprise Information Program, Priority Ecosystems Science, the Fort Collins Science Center, and the Central Region Geospatial Information Office

Scientific Investigations Report 2007–5232

U.S. Department of the Interior
U.S. Geological Survey

U.S. Department of the Interior
DIRK KEMPTHORNE, Secretary

U.S. Geological Survey
Mark D. Myers, Director

U.S. Geological Survey, Reston, Virginia 2007

For product and ordering information:
World Wide Web: http://www.usgs.gov/pubprod
Telephone: 1-888-ASK-USGS

For more information on the USGS—the Federal source for science about the Earth, its natural and living resources, natural hazards, and the environment:
World Wide Web: http://www.usgs.gov
Telephone: 1-888-ASK-USGS

Cover: Lake and Mountains Scenic View (Alaska).
Photograph by John J. Mosesso/NBII Image Gallery.

Suggested citation:
Henkel, Heather S., 2007, Proceedings of the first U.S. Geological Survey scientific information management workshop, March 21–23, 2006: U.S. Geological Survey Scientific Investigations Report 2007-5232, 94 p.

Contents

Abbreviations Used in This Report	
ARCIMS	Arc Internet Map Server
BRD	Biological Resources Discipline (USGS)
CD-R	Compact Disc Recordable
CD-ROM	Compact Disc Read-Only Memory
CMGP	Coastal and Marine Geology Program (USGS)
COI	Communities of Interest
CoP	Communities of Practice
DAC	Data Advisory Committee (DOI)
ESRI	Environmental Systems Research Institute
FGDC	Federal Geographic Data Committee
GD	Geology Discipline (USGS)
GIO	Geospatial Information Office (USGS)
GIS	Geographic Information System
GODM	Global Organism Detection and Monitoring System (NIISS)
IPANE	Invasive Plant Atlas of New England
IWGBSC	Interagency Working Group on Scientific Collections
KOS	Knowledge Organization System
LASED	Louisiana Sedimentary and Environmental Database (USGS)
LIDAR	Light Detection and Ranging
LTER	Long Term Ecological Research Network
MRIB	Marine Realms Information Bank (USGS)
NARA	National Archives and Records Administration
NASA	National Aeronautics and Space Administration
NAWQA	National Water-Quality Assessment (USGS)
NBII	National Biological Information Infrastructure
NGDB	National Geochemical Database (USGS)
NIH	National Institutes of Health
NIISS	National Institute of Invasive Species Science
NOAA	National Oceanic and Atmospheric Administration
NRC	National Research Council
OSTP	Office of Science and Technology Policy
REST	Representational State Transfer
SIM	Scientific Information Management
SOAP	Simple Object Access Protocol
SODA	Self-service Online Digital Archive
SOFIA	South Florida Information Access (USGS)
USDA	U.S. Department of Agriculture
USDOI	U.S. Department of the Interior
USEPA	U.S. Environmental Protection Agency
USFWS	U.S. Fish and Wildlife Service
USGS	U.S. Geological Survey

I. Introduction

Abstract

In March 2006, the U.S. Geological Survey (USGS) held the first Scientific Information Management (SIM) Workshop in Reston, Virginia. The workshop brought together more than 150 SIM professionals from across the organization to discuss the range and importance of SIM problems, identify common challenges and solutions, and investigate the use and value of "communities of practice" (CoP) as mechanisms to address these issues.

The 3-day workshop began with presentations of SIM challenges faced by the Long Term Ecological Research (LTER) network and two USGS programs from geology and hydrology. These presentations were followed by a keynote address and discussion of CoP by Dr. Etienne Wenger, a pioneer and leading expert in CoP, who defined them as "groups of people who share a passion for something that they know how to do and who interact regularly to learn how to do it better." Wenger addressed the roles and characteristics of CoP, how they complement formal organizational structures, and how they can be fostered. Following this motivating overview, five panelists (including Dr. Wenger) with CoP experience in different institutional settings provided their perspectives and lessons learned. The first day closed with an open discussion on the potential intersection of SIM at the USGS with SIM challenges and the potential for CoP.

The second session began the process of developing a common vocabulary for both scientific data management and CoP, and a list of eight guiding principles for information management were proposed for discussion and constructive criticism. Following this discussion, 20 live demonstrations and posters of SIM tools developed by various USGS programs and projects were presented.

Two community-building sessions were held to explore the next steps in 12 specific areas: Archiving of Scientific Data and Information; Database Networks; Digital Libraries; Emerging Workforce; Field Data for Small Research Projects; Knowledge Capture; Knowledge Organization Systems and Controlled Vocabularies; Large Time Series Data Sets; Metadata; Portals and Frameworks; Preservation of Physical Collections; and Scientific Data from Monitoring Programs. In about two-thirds of these areas, initial steps to forming CoP are now underway.

The final afternoon included a panel in which information professionals, managers, program coordinators, and associate directors shared their perspectives on the workshop, on ways in which the USGS could better manage its scientific information, and on the use of CoP as informal mechanisms to complement formal organizational structures. The final session focused on developing the next steps, an action plan, and a communication strategy to ensure continued development.

— Thomas Gunther

Executive Summary

The U.S. Geological Survey (USGS) Scientific Information Management (SIM) Workshop brought together more than 150 participants to learn ways to better manage scientific information. The workshop was held March 21–23, 2006, at USGS headquarters in Reston, Virginia. The USGS SIM Workshop was co-sponsored by the Coastal and Marine Geology Program, the Enterprise Information Program, the Priority Ecosystems Science Program, the Fort Collins Science Center, and the Central Region Geospatial Information Office. Discussions and presentations centered on crosscutting issues, common problems, and the value of communities of practice (CoP) as mechanisms to collaborate with one another to better manage scientific information at USGS.

Thomas Gunther, SIM Workshop Chairperson opened the meeting by reviewing the objectives, thanking the co-sponsors, and introducing Dr. P. Patrick Leahy, Acting Director, USGS. Dr. Leahy began with an overview of the importance of information management at the USGS. He referenced a 2001 report from the National Research Council (NRC), which stated that "one of the USGS' most valuable assets is its long-term data sets," and he said that better management of our information would help achieve the vision of the USGS as a "natural science and information agency."[1]

Dr. Leahy pointed out that there are at least three other benefits from good SIM practices: more effective use of USGS resources; a stronger science organization, because data and information are more readily available to researchers; and more valuable science products, because they can be more easily found and used. Basically, the preservation of scientific information for future researchers is a fundamental obligation of scientists and scientific organizations. Dr. Leahy also suggested that we find ways to learn from each other, collaborate in the development of information-management tools and strategies, and use ideas such as CoP to take advantage of the skills and knowledge that already exist in the Bureau. Finally, he thanked the workshop sponsors and organizers and said that he looked forward to the results of the workshop.

The morning program, hosted by Ronnie Best, continued with presentations by scientists and information managers on the diversity of challenges and approaches to information and data management. Key ideas focused on the integration of tools and data sets, the importance of metadata, the power of using volunteers, and how to present our data more effectively for a broader audience.

The first afternoon focused on CoP as mechanisms to better manage our scientific information. Dr. Etienne Wenger, the keynote speaker and a pioneer and leading expert in CoP, defined CoP as "groups of people who share a passion for something that they know how to do and who interact regularly to learn how to do it better." He noted the community is a complement to the individual and that practitioners need a community to help solve problems, validate ideas, push the boundaries of their fields, and think of new

[1] *Future Roles and Opportunities for the U.S. Geological Survey*, National Research Council, 2001.

ways to leverage knowledge. Throughout his presentation Dr. Wenger stressed the importance of the success factors that include participation, nurturing, sponsorship, and support. Participants learned about the foundational elements of CoP—domain, community, and practice. Next, they discovered, through small group discussions guided by Dr. Wenger, that CoP are really nothing new and that we have all interacted with people who have a similar interest or common professional practice. In closing, he pointed out that organizations are too complex to expect formal structures to manage everything and that reorganizations cannot address every need. Informal structures such as CoP can complement and connect with the formal organization.

Dr. Wenger moderated the panel discussion that followed. Panel members from the U.S. Fish and Wildlife Service (USFWS), ICF Consulting, U.S. Forest Service, and USGS Office of Human Resources discussed the CoP approach as applied within their organizations. There was a substantial amount of interaction as different perspectives emerged among the audience, panel members, and moderator.

The closing discussion, moderated by Dr. Ronnie Best, served to summarize the activities of the first day while posing a number of questions and issues for further exploration. These issues included the differences between a community and a committee; the role of management, sponsors, and champions; ensuring that the community has a voice in decisions; and recognizing that different models may be used.

Bill Miller opened the second day by engaging participants in an activity and discussion focused on the development of common definitions for information- and data-management terminology. This activity led to scheduled demonstrations and a poster session organized by Susan Russell-Robinson. Attendees spent the remainder of the morning viewing posters and demonstrations and networking with colleagues.

After lunch, David Govoni set the context for the community-building sessions. Participants chose one of six community-building sessions in which to participate for the afternoon. Sessions on Archiving, Digital Libraries, Small Research Projects, Knowledge Capture, Portals and Frameworks, and Monitoring Programs were convened to discuss scope, issues, needs, and level of interest. Discussion leaders then reported key decisions, recommendations, and next steps at the closing plenary session.

Sky Bristol provided the overview for Thursday. A second set of community-building sessions gave attendees an opportunity to participate in and help frame six more CoP. Focus areas were Database Networks, Emerging Workforce, Knowledge Organization Systems and Controlled Vocabularies, Large Time Series Data Sets, Metadata, and Preservation of Physical Collections. As a result of the 12 community-building sessions, a number of CoP are now emerging, and collaborative tools are being identified and developed.

Martha Garcia moderated the afternoon panel in which information professionals, managers, program coordinators, and associate directors shared their perspectives on the workshop, on ways in which the USGS could better manage its scientific information,

and on the use of CoP as informal mechanisms to complement formal organizational structures. The final session focused on developing the next steps, an action plan, and a communication strategy to ensure continued development.

Tom Gunther closed the final session by thanking those who were involved in making the workshop a success and for launching the USGS into a new frontier for managing its scientific information.

Workshop Objectives

- Compile initial listing of needs, tools, and best practices for SIM challenges that are, or can be made available to USGS programs and projects.
- Recognize existing SIM groups or "CoP" in the USGS and recommend actions regarding the community framework for addressing additional SIM challenges.
- Agree upon methods for scientific information managers to communicate with each other; exchange knowledge, tools and approaches; and develop collaborative efforts.
- Identify a needed suite of tools to facilitate communication within and among groups or communities.

Summary of Remarks

Day 1, Morning Session (March 21, 2006)

Welcome to the workshop – Thomas Gunther, SIM Workshop Chair

- Overview of workshop objectives and outcomes
- Introduce Workshop Host Committee and Workshop Sponsors—the Coastal and Marine Geology Program, the Enterprise Information Program, the Priority Ecosystems Science Program, the Fort Collins Science Center, and the Central Region Geospatial Information Office
- Introduce Facilitation Team – Tricia Gibbons, Kathleen Cleary, and Rodney Payne

Introduction of the Acting Director – Thomas Gunther

Director's Charge for the Workshop – Dr. P. Patrick Leahy
Highlights:

- Welcome and pleased to have a meeting that is so important.
- Workshops are critical mechanisms by which an organization learns.
- Scientists “owned” their data in the past. This changed in the past decade as we realized our information has value. Challenge that there be no “unloved data,” i.e., data that are available but unused or under used.
- Lessons Learned from Katrina – Why information that was possessed was not used. We cannot allow this to happen in the future. Information Management is critical. We need to package our information better for broader markets—in a sense, market this asset better.

- We are a scientific INFORMATION agency. That is why this workshop is so important.
- The three reasons are limited resources, people, and accountability.
- We need to collaborate better—efficiently and effectively.
- We need to build a stronger science organization.
- We need to pay more attention to self-organized teams. There is a lot of power in this. We need to question ourselves so we don't get stale.
- In the Gulf Coast, USGS must be a player. We provide science and information for the tough decisions.
- Thank you to the workshop sponsors.

Science Information Management: Practices, Challenges, and Opportunities

This session, hosted by Ronnie Best, was a forum for presentations and questions by natural and information scientists and information managers on the diversity of approaches to information and data management. Presentations included

- Data-Management Challenges for the USGS Volcano Hazards Program—Dr. Peter Cervelli and Dr. Jim Quick
- Things that LTER Learned Managing Long-Term Data Sets—Dr. Indigo San Gil
- The IPANE Program: The Synergism of Science and Public Involvement—Dr. Leslie J. Mehrhoff
- Making Sense of it all: An Ecologist's Perspective on National Databases and Data Analysis Tools in the NAWQA Program—Dr. Thomas F. Cuffney

Key Themes and Ideas from Day 1, Morning Session

- Integration of tools and data analysis
- Shared uniform data standards
- Demonstrate the importance of metadata—worth the effort
- Archiving
- Tools repositories
- Power of using volunteers
- More effectively serve up/present our data for a broader audience, end-users
- New ways of learning from one another—sharing ideas and data
- Building the infrastructure up front
- Difficulty of long-term funding
- Giving data life beyond the projects
- What about analytical tools
- Value of science data
- Respond quickly and effectively
- Learn from ours and other peoples' mistakes and successes
- Need to capture and share lessons learned and best practices

Day 1, Afternoon Session

Introduction of Keynote Speaker – Thomas Gunther

Keynote Address - Dr. Etienne Wenger, Learning for a Small Planet

Dr. Wenger is a leading expert on CoP. He is the founder of Learning for a Small Planet, an investigation into fostering learning institutions. He also is a former Research Scientist at the Institute for Research on Learning, where he developed his learning theory centered on the concept of CoP. For the last 6 years, he has been helping organizations develop and implement knowledge strategies based on CoP.

Highlights:

- CoP are really nothing new—learning and interaction with people that have a similar interest or common professional practice.
- The community is a complement to the individual.
- Foundational elements—domain, community, and practice
- Success factors—participation, nurturing, sponsorship, and support
- Practitioners need a community to help solve problems, validate ideas, push the boundaries of their fields, and think of new ways to leverage knowledge.
- Organizations are too complex for our formal structures to manage everything.
- We must manage the boundary between the formal and informal structures, and let the informal complement and connect with the formal.
- Breakouts discussion debrief
    - There is a delicate balance between sponsorship and support.
    - Communities are not good places to manage jobs or firings.
    - Communities have projects—that's where sponsorship comes in—to open the door for some resources.
    - It is not a good practice to assign people to a community.
    - Energizing tasks and de-energizing tasks; very important to make the difference.
    - Some communities are hard to kill—if you're going to dance with a bear, you can't decide when to stop.
    - Need a commitment to the community—a goal is critical but it can be a broad goal of learning together (open-ended goals).
    - Formal and Informal—the future is more in the informal; constant dance between the formal and informal.
- Seems like some free time is needed—time commitment has to be recognized as work time.
- CoP must have sponsorship.

- Most communities have an online component. New technology has changed views. It takes more care—you must say something to create a sense of community. Rhythm also is needed.
- Teleconference really can be good.

Panel I — Communities *in* Practice

Panel members discussed the CoP approach as applied and practiced in other organizations as well as the USGS. Panel members and highlights included

- Dr. Etienne Wenger (Convener)
- Bill Knapp (USFWS)
  - Knowledge management
  - Succession planning
  - Commitment and support
  - Need a few champions at the high level of an organization for it to work—vertical support is necessary.
- Mark Youman (ICF Consulting)
  - Think in a different context of organization—outside the organization. Federal Highway Administration is a good example.
  - What is the right model for our community? Formal and Informal, internal and external, expectations and contributions
  - What are the ways to measure community? It is important to have objectives and measures and to re-visit them.
- Laure Wallace (USGS Office of Human Resources)
  - Examples of CoP —EROS Data Center, Denver, and Reston Leadership groups
  - Leadership groups share ideas and practice without the title.
  - Culture change is necessary for CoP to work.
- Susan Mockenhaupt (U.S. Forest Service)
  - Work with communities of **passion** along with CoP.
  - CoP can be very counter culture.
  - Need some tools to be successful.

Open Discussion — The Intersection of CoP and SIM at the USGS

Ronnie Best moderated this dialog among the morning presenters, the keynote speaker, the panel members, and the audience. Key discussion points included

- How do you see this practice useful in your role? As scientists, we have been trained to do this, but we need managers willing to take this on and support the concept.
- How do you get management to change? Instead of trying to change management, just work with them.

- How much involvement do we have or need from the managers?
- CoP need to engage the end-users. End-users are a very large and important group.
- SIM—three most important things are
    - Archive data—how do you do it?
    - Metadata issues—capture and storage
    - Making data available
- Work with each other so each can understand what you are doing and why.
- USGS better connected to understanding USFWS.
- USGS needs to help USFWS remember what a good science organization looks like.
- Forestry needs mapping from USGS.
- Confused—Are we talking about committees or CoP? How are they different in reality? Valid point.
- Committee versus Community. Are they open and self-selecting? Are they closed and only by invitation? A: Different models out there—better to start open. Consider communities of place and interest.
- What is the role for the sponsor? Role of the sponsor is important.
- Community has to feel respected. Ownership is crucial for success.
- Choice is VERY important. A VOICE is important. The community does not have to make the decision but must feel it has a VOICE in the decision.
- How are the end-users using our science? Need to ask these questions. How do we deliver this science?
- Ways to kill off CoP
    - Too much structure
    - Conditions to be a member will "kill the goose before the egg is laid."
    - "Lighten Up"
- Need different models
- Working together without scaring people away

Day 2, Morning Session (March 22, 2006)

Committee Host: Bill Miller

Key Lessons:
Where we left off yesterday and where we are going today

- People not in CoP really don't understand them.
- If you want to begin a community of practice, just do it. You don't need a lot of approvals from higher up.
- CoP should not be dictated from above.
- The inherent benefits of a community of practice are to the individual. It could have a benefit for others and to the organization but not necessarily.
- Ideas with merit will succeed.
- CoP don't need a definite task or goal but the sharing of ideas could be the goal itself.
- Everyone sees CoP from different perspectives.
- Consider communities of place and communities of interest.
- Knowledge and learning communities
- Common understanding—share and build from what is already known.
- Respect for the community
- Owning your own domain—meaning you have a voice in the decision, not necessarily the choice.
- Recognize that different models can be operating with different levels of formality.
- We need to make our data more available and understandable to those outside the USGS. Impact of the end-user must be considered.

Group Activity — Words Matter: Developing a Common Vocabulary for Common Understanding

Bill Miller introduced a group activity for reviewing definitions of concepts, standards, and processes to arrive at common terminology for information and data management. After presenting the eight principles (listed in Appendix II), participants had an opportunity to discuss, share, and reach a consensus on the definition of "data." An interesting and lively discussion followed.

Open Session — Demonstrations and Posters

This session was organized and coordinated by Susan Russell-Robinson. The demonstrations and posters focused on information-management issues as well as tools and systems used for information and data management that may be shared among programs. Live demonstrations were held according to schedule in the Visitor's Center and the Fine Arts Hallway. Poster sessions were held continuously in the Fine Arts Hallway and on the stage of the auditorium during the morning session. Contributed abstracts from the demonstrations and exhibits are available in Section V.

Live Demonstrations

- "Geographic Information System for the Gulf – ADS40 Imagery on LIDAR for Hurricane Katrina" – David Greenlee
- "Data Reference Model 2.0 and Metadata" – Danielle Forsyth
- "Database of Geology Databases" – Jerry McFaul
- "Geospatial One-Stop Community Geographic Information System Portal Application" – Steven Hale and Marc Levine
- "MRIB – A Digital Library for Coastal and Marine Science" – Alan Allwardt and Fran Lightsom
- "myUSGS Portal Pilot Project" – Sky Bristol
- "StreamStats Web Application" – Kernell Reis
- "National Water-Quality Assessment Program Invertebrate and Algal Data Analysis Software" – Thomas F. Cuffney
- "Global Invasive Species Information Network" – Catherine Jarnevich
- "Return on Investment on Web Metrics at USGS" – Kit Fuller
- "USGS Science Topics Index and Supporting Infrastructure" – Peter Schweitzer

Poster Sessions

- "Five Federal Information Programs" – Carm Ferrigno, John Faundeen and Carol Wippich
    1. Records Management
    2. Privacy Act
    3. Freedom of Information Act
    4. Capital Planning and Investment Control
    5. Section 508 Accessibility

- "Formalizing the Information Product Lifecycle" – Wendy Danchuk
- "USGS Publications Warehouse" – Greg Allord and Wendy Danchuk
- "South Florida Information Access (SOFIA)" – Heather Henkel
- "Utilization of the Data Reference Model for Improved Data Sharing" – Ray Obuch, Stu Doescher and Tony Frank
- "The Geodatabase Solution to Data Management: Examples from LASED and XSTORMS" – Shawn Dadisman
- "Collaboration in the Development of the Bird Banding Lab Database" – Kevin Laurent
- "Vector Data Themes" – Jerry Ornelas
- "Development and Utilization of the National Geochemical Database" – Steve Smith

Day 2, Afternoon Session

Committee Host: David Govoni
David Govoni opened the afternoon session by providing an overview of the community strategy and charge to communities.

What are the Differences Between Communities and Teams?

TEAMS	COMMUNITIES
• Driven by deliverables • Defined by task • Develop by work plan • Bound by commitment	• Driven by value • Defined by knowledge • Develop through learning • Bound by identity

Objectives and responsibilities of the communities were reviewed before participants moved to the breakout sessions.

Community-Building Sessions I
Participants chose to take part in one of the six community sessions for the afternoon. Groups reported out key decisions, recommendations, or next steps resulting from the community work session. Comments from some of these sessions are available in Appendix III.

Day 3, Morning Session (March 23, 2006)

Committee Host – Sky Bristol

- Overview of Day 3 activities and setup for CoP sessions – Rodney Payne
- Recap of Day 2 community-building sessions – Sky Bristol
    - Collaboration tools: list server, usgs.gov, e-mail forum, online forum, Wiki
    - Beginning to network—great outcome of the workshop
    - Take this network and figure out how to use it
    - Idea: personal Wiki, staff member is assigned his/her own page

Community-Building Sessions II
Participants chose to take part in one of the six community sessions for the morning. Groups reported out key decisions, recommendations, or next steps resulting from the community work session. Comments from some of these sessions are available in Appendix III.

Day 3, Afternoon Session

Participants were asked what they liked and disliked about the meeting. Their comments are captured below.

Like About the Meeting	Suggested Changes
Networking with other people with common interests Learning about other tools Kick starting communities Learning about recent legislation Having the real leaders in the field HERE to interact with us How other agencies are using CoP Lots of enthusiasm to keep going Dr. Leahy came down to open the workshop Working together to solve our issues around science information management "I even talked with a biologist"	Wish we had been able to go to more than two sessions Registration process was confusing Some confusion with demos and posters Expand telecom capabilities Like to see more researchers as participants Bogged down late in the first day (last panel not needed) Breakout groups should start earlier in the program to make connections

Panel II — Perspectives on the Workshop – Martha Garcia, Convener

Panel members shared perspectives, impressions, and issues related to SIM and implementation of CoP. The panel included

- Martha Garcia (USGS-BRD Priority Ecosystems Science)
- Mark DeMulder (GIO Science Information and Education Office)
- John Faundeen (USGS Archivist)
- Linda Gunderson (USGS-GD Acting Associate Director for Geology)
- Fran Lightsom (USGS-GD Coastal and Marine Geology Program)
- Peter Lyttle (USGS-GD Cooperative Geologic Mapping Program)

Highlights and Perspectives:

- Heard people feel like they are isolated; this is a good networking opportunity.
- Tools are being developed; for example, establishing Wikis.
- Need to take advantage of the technology that is out there.
- There is passion about improving how we manage scientific information.
- Retirees are taking knowledge and data with them—legacy data problem.
- Metadata problems still exist—need to be part of planning and culture.
- NARA role not understood.
- Access and preservation to data is still a concern.
- Need to be sharing tools.

- Need to truly manage all of the data—mixture of the informal and formal.
- Need the freedom of the communities.
- Managers need to create the environment for CoP to emerge.
- What's the role of the GIO in supporting this?
- Senior management needs to understand the management of the data.
- Thanks to the staff that made this happen; this was a very meaningful workshop.
- Heard some confusion about the GIO.
- The Invasive Plant Atlas of New England (IPANE) success is built on volunteers; we need more of this.
- Heard five things from the breakouts: database of database needed, need for collaboration tools, the difficulties dealing with legacy data sets, the need to geospatially enable digital holdings, and turning data into information and information into knowledge.
- The CoP might be one tool for information management.
- CoP are essential to all of us.
- Beginning to create an environment for information management.
- We are socializing our learning and knowledge by sharing.
- Money and management support are needed; Linda Gunderson will commit to financial support of CoP.
- CoP are ways of making science information at USGS a resilient process.
    - Empowerment through sharing information and best practices
    - Constant communication for 3 days must continue, even the bad news.
    - Can't be captivated by the tools; keep the big picture in front of us.

Workshop Closing and Adjourn

Tom Gunther thanked participants, workshop sponsors, the host committee, session leaders, and the facilitation team for contributing to a very successful workshop.

II. Discussion Papers

1. "Communities of Practice, A Brief Introduction" – Dr. Etienne Wenger, Learning for a Small Planet
2. "Communities of Interest and Communities of Practice: Components of a Larger Scientific Enterprise" – William G. Miller, USGS
3. "Knowledge Capture" – Laure Wallace, USGS
4. "SODA—A Self-service Online Digital Archive" – Rex Sanders, USGS

Communities of Practice, A Brief Introduction

Etienne Wenger, Learning for a Small Planet

The term "community of practice" is of relatively recent coinage, even though the phenomenon it refers to is age-old. The concept has turned out to provide a useful perspective on knowing and learning. A growing number of people and organizations in various sectors are now focusing on communities of practice as a key to improving their performance.

This brief and general introduction examines what communities of practice are and why researchers and practitioners in so many different contexts find them useful as an approach to knowing and learning.

What are communities of practice?

Communities of practice are formed by people who engage in a process of collective learning in a shared domain of human endeavor: a tribe learning to survive, a band of artists seeking new forms of expression, a group of engineers working on similar problems, a clique of pupils defining their identity in the school, a network of surgeons exploring novel techniques, a gathering of first-time managers helping each other cope. In a nutshell:

> Communities of practice are groups of people who share a concern or a passion for something they do and learn how to do it better as they interact regularly.

Note that this definition allows for, but does not assume, intentionality: learning can be the reason the community comes together or an incidental outcome of member's interactions. Not everything called a community is a community of practice. Not everything called a community is a community of practice. A neighborhood for instance, is often called a community, but is usually not a community of practice. Three characteristics are crucial:

1. **The domain:** A community of practice is not merely a club of friends or a network of connections between people. It has an identity defined by a shared domain of interest. Membership therefore implies a commitment to the domain, and therefore a shared competence that distinguishes members from other people. (You could belong to the same network as someone and never know it.) The domain is not necessarily something recognized as .expertise. outside the community. A youth gang may have developed all sorts of ways of dealing with their domain: surviving on the street and maintaining competence and learn from each other, even though few people outside the group may value or even recognize their expertise.

2. **The community:** In pursuing their interest in their domain, members engage in joint activities and discussions, help each other, and share information. They build relationships that enable them to learn from each other. A website in itself is not a community of practice. Having the same job or the same title does not make for a community of practice unless members interact and learn together. The claims processors in a large insurance company or students in American high schools may have much in common, yet unless they interact and learn together, they do not form a community of practice. But members of a community of practice do not necessarily work together on a daily basis. The Impressionists, for instance, used to meet in cafes and studios to discuss the style of painting they were inventing together. These interactions were essential to making them a community of practice even though they often painted alone.

3. **The practice:** A community of practice is not merely a community of interest--people who like certain kinds of movies, for instance. Members of a community of practice are practitioners. They develop a shared repertoire of resources: experiences, stories, tools, ways of addressing recurring problems.in short a shared practice. This takes time and sustained interaction. A good conversation with a stranger on an airplane may give you all sorts of interesting insights, but it does not in itself make for a community of practice. The development of a shared practice may be more or less self-conscious. The .windshield wipers. engineers at an auto manufacturer make a concerted effort to collect and document the tricks and lessons they have learned into a knowledge base. By contrast, nurses who meet regularly for lunch in a hospital cafeteria may not realize that their lunch discussions are one of their main sources of knowledge about how to care for patients. Still, in the course of all these conversations, they have developed a set of stories and cases that have become a shared repertoire for their practice.

It is the combination of these three elements that constitutes a community of practice. And it is by developing these three elements in parallel that one cultivates such a community.

What do communities of practice look like?

Communities develop their practice through a variety of activities. The following table provides a few typical examples:

Problem solving	"Can we work on this design and brainstorm some ideas; I'm stuck."
Requests for Information	"Where can I find the code to connect to the server?"
Seeking experience	"Has anyone dealt with a customer in this situation?"
Reusing assets	"I have a proposal for a local area network I wrote for a client last year. I can send it to you and you can easily tweak it for this new client."
Coordination and	"Can we combine our purchases of solvent to achieve bulk

synergy	discounts?"
Discussing developments	"What do you think of the new CAD system? Does it really help?"
Documentation projects	"We have faced this problem five times now. Let us write it down once and for all."
Visits	"Can we come and see your after-school program? We need to establish one in our city."
Mapping knowledge and identifying gaps	"Who knows what, and what are we missing? What other groups should we connect with?"

Communities of practice are not called that in all organizations. They are known under various names, such as learning networks, thematic groups, or tech clubs.

While they all have the three elements of a domain, a community, and a practice, they come in a variety of forms. Some are quite small; some are very large, often with a core group and many peripheral members. Some are local and some cover the globe. Some meet mainly face-to-face, some mostly online. Some are within an organization and some include members from various organizations. Some are formally recognized, often supported with a budget; and some are completely informal and even invisible.

Communities of practice have been around for as long as human beings have learned together. At home, at work, at school, in our hobbies, we all belong to communities of practice, a number of them usually. In some we are core members. In many we are merely peripheral. And we travel through numerous communities over the course of our lives.

In fact, communities of practice are everywhere. They are a familiar experience, so familiar perhaps that it often escapes our attention. Yet when it is given a name and brought into focus, it becomes a perspective that can help us understand our world better. In particular, it allows us to see past more obvious formal structures such as organizations, classrooms, or nations, and perceive the structures defined by engagement in practice and the informal learning that comes with it.

Where does the concept come from?

Social scientists have used versions of the concept of community of practice for a variety of analytical purposes, but the origin and primary use of the concept has been in learning theory. Anthropologist Jean Lave and I coined the term while studying apprenticeship as a learning model. People usually think of apprenticeship as a relationship between a student and a master, but studies of apprenticeship reveal a more complex set of social relationships through which learning takes place mostly with journeymen and more advanced apprentices. The term community of practice was coined to refer to the community that acts as a living curriculum for the apprentice. Once the concept was articulated, we started to see these communities everywhere, even when no formal apprenticeship system existed. And of course, learning in a community of practice is not

limited to novices. The practice of a community is dynamic and involves learning on the part of everyone.

Where is the concept being applied?

The concept of community of practice has found a number of practical applications in business, organizational design, government, education, professional associations, development projects, and civic life.

Organizations. The concept has been adopted most readily by people in business because of the recognition that knowledge is a critical asset that needs to be managed strategically. Initial efforts at managing knowledge had focused on information systems with disappointing results. Communities of practice provided a new approach, which focused on people and on the social structures that enable them to learn with and from each other. Today, there is hardly any organization of a reasonable size that does not have some form communities-of-practice initiative. A number of characteristics explain this rush of interest in communities of practice as a vehicle for developing strategic capabilities in organizations:

- Communities of practice enable practitioners to take collective responsibility for managing the knowledge they need, recognizing that, given the proper structure, they are in the best position to do this.
- Communities among practitioners create a direct link between learning and performance, because the same people participate in communities of practice and in teams and business units.
- Practitioners can address the tacit and dynamic aspects of knowledge creation and sharing, as well as the more explicit aspects.
- Communities are not limited by formal structures: they create connections among people across organizational and geographic boundaries.

From this perspective, the knowledge of an organization lives in a constellation of communities of practice each taking care of a specific aspect of the competence that the organization needs. However, the very characteristics that make communities of practice a good fit for stewarding knowledge - autonomy, practitioner-orientation, informality, crossing boundaries - are also characteristics that make them a challenge for traditional hierarchical organizations. How this challenge is going to affect these organizations remains to be seen.

Government. Like businesses, government organizations face knowledge challenges of increasing complexity and scale. They have adopted communities of practice for much the same reasons, though the formality of the bureaucracy can come in the way of open knowledge sharing. Beyond internal communities, there are typical government problems such as education, health, and security that require coordination and knowledge sharing across levels of government. There also, communities of practice hold the promise of enabling connections among people across formal structures. And there also, there are substantial organizational issues to overcome.

Education. Schools and districts are organizations in their own right, and they too face increasing knowledge challenges. The first applications of communities of practice have been in teacher training and in providing isolated administrators with access to colleagues. There is a wave of interest in these peer-to-peer professional-development activities. But in the education sector, learning is not only a means to an end: it the end product. The perspective of communities of practice is therefore also relevant at this level. In business, focusing on communities of practice adds a layer of complexity to the organization, but it does not fundamentally change what the business is about. In schools, changing the learning theory is a much deeper transformation. This will inevitably take longer. The perspective of communities of practice affects educational practices along three dimensions:

- *Internally:* How to organize educational experiences that ground school learning in practice through participation in communities around subject matters?
- *Externally:* How to connect the experience of students to actual practice through peripheral forms of participation in broader communities beyond the walls of the school?
- *Over the lifetime of students:* How to serve the lifelong learning needs of students by organizing communities of practice focused on topics of continuing interest to students beyond the initial schooling period?

From this perspective, the school is not the privileged locus of learning. It is not a self-contained, closed world in which students acquire knowledge to be applied outside, but a part of a broader learning system. The class is not the primary learning event. It is life itself that is the main learning event. Schools, classrooms, and training sessions still have a role to play in this vision, but they have to be in the service of the learning that happens in the world.

Associations. A growing number of associations, professional and otherwise, are seeking ways to focus on learning through reflection on practice. Their members are restless and their allegiance is fragile. They need to offer high-value learning activities. The peer-to-peer learning activities typical of communities of practice offer a complementary alternative to more traditional course offerings and publications.

Social sector. In the civic domain, there is an emergent interest in building communities among practitioners. In the non-profit world, for instance, foundations are recognizing that philanthropy needs focus on learning systems in order to fully leverage funded projects. But practitioners are seeking peer-to-peer connections and learning opportunities with or without the support of institutions. This includes regional economic development, with intra-regional communities on various domains, as well as inter-regional learning with communities gathering practitioners from various regions.

International development. There is increasing recognition that the challenge of developing nations is as much a knowledge as a financial challenge. A number of people believe that a communities-of-practice approach can provide a new paradigm for

development work. It emphasizes knowledge building among practitioners. Some development agencies now see their role as conveners of such communities, rather than as providers of knowledge.

The web. New technologies such as the Internet have extended the reach of our interactions beyond the geographical limitations of traditional communities, but the increase in flow of information does not obviate the need for community. In fact, it expands the possibilities for community and calls for new kinds of communities based on shared practice.

The concept of community of practice is influencing theory and practice in many domains. From humble beginnings in apprenticeship studies, the concept was grabbed by businesses interested in knowledge management and has progressively found its way into other sectors. It has now become the foundation of a perspective on knowing and learning that informs efforts to create learning systems in various sectors and at various levels of scale, from local communities, to single organizations, partnerships, cities, regions, and the entire world.

Further reading

For the application of a community-based approach to knowledge in organizations

- Cultivating communities of practice: a guide to managing knowledge. By Etienne Wenger, Richard McDermott, and William Snyder, Harvard Business School Press, 2002.
- Communities of practice: the organizational frontier. By Etienne Wenger and William Snyder. Harvard Business Review. January-February 2000, p. 139–145.
- Knowledge management is a donut: shaping your knowledge strategy with communities of practice. By Etienne Wenger. Ivey Business Journal, January 2004.

For technology issues:

- Supporting communities of practice: a survey of community-oriented technologies. By Etienne Wenger. Self-published report available at http://www.ewenger.com/tech, 2001.

For in-depth coverage of the learning theory:

- Communities of practice: learning, meaning, and identity. By Etienne Wenger, Cambridge University Press, 1998.

For a vision of where the learning theory is going:

- Learning for a small planet: a research agenda. By Etienne Wenger, available at http://www.ewenger.com/research, 2004.

Communities of Interest and Communities of Practice: Components of a Larger Scientific Enterprise

William G. Miller
U.S. Geological Survey

Abstract

What are the differences, if any, between communities of interest and communities of practice in the scientific enterprise? Communities of interest (COI) are groups formed to create a marketplace for the intellectual property created using the scientific method. These communities may have a rigid structure and limited membership. Communities of practice (CoP) are self-associating groups specifically organized as a learning tool for their membership. While related, these groups have very different goals and methods.

Introduction

Why should one be interested in communities, let alone specialized types of communities? The answer has two components. First, the process of conducting scientific research is an inherently social activity (De Mey, 1982; Kuhn, 1970). Second, communities can form an environment for learning while performing work and sharing knowledge (Wenger and others, 2002). Knowing the different types of communities that form in the scientific enterprise can lead directly to personal and organizational growth and reward.

While very interested in this subject, the author is neither a sociologist nor a psychologist. The author must therefore, in addition to personal observations, make use of the work of those who are. The purpose of this paper is to investigate COI and CoP, trying to understand why they occur and what assistance they are to the people who join them.

It will be argued that there are several types of COI, each with different rules and organizational styles. It also will be argued that the CoP are a subset of COI.

The first step in understanding these complex social entities is to try to differentiate between the two. The first problem is that many people are not precise when they speak of communities. Four such descriptions are

> A community of interest (COI) "is a term used to describe any collaborative group of users who must exchange information in pursuit of their shared goals, interests, missions, or business processes, and who therefore must have shared vocabulary for the information they exchange. The COI concept is very broad, and covers an enormous number of potential groups of every kind and size" (Department of Defense, 2004, p. 3).

> "Communities of interest are groups of people (typically coming from different disciplines) which engage in a joint activity. The "symmetry of ignorance" or "asymmetry of knowledge" among the different stakeholders within a communities [sic] of interest serves as a challenge to create new knowledge and shared understanding" (enTWIne Project, 2002).

> A community of practice (CoP) is "a flexible group of professionals, informally bound by common interests, who interact through interdependent tasks guided by a common purpose thereby embodying a store of common knowledge" (Jubert, 1999, p. 166).

> "Communities of practice steward the knowledge assets of organizations and society. They operate as "social learning systems" where practitioners connect to solve problems, share ideas, set standards, build tools, and develop relationships with peers and stakeholders. These structures are considered informal because they cannot be mandated from the outside. An essential dimension of a community of practice is voluntary participation, because without this a member is less likely to seek or share knowledge; build trust and reciprocity with others; or apply the community's knowledge in practice" (Snyder and Briggs, 2003, p. 7).

Each of these definitions, with the possible exception of Snyder and Briggs', seem to be describing essentially the same activity.

Communities of Interest

COI are not new. The scientific method that we know today is based on one concept of "COI." Sir Francis Bacon (1561–1626) (Encyclopedia Britannica, 2006a) was instrumental in developing the transition from the Aristotelian view of

science[2] to the view that we have today. His argument began with the publication of *The Advancement of Learning* (Bacon, 1605), followed by *Novum Organum* [new organon] (Bacon, 1620), and *The New Atlantis* (Bacon, 1626). Paolo Rossi has observed:

> "Bacon proposed to the European culture an alternative view of science. For him science had a public, democratic, and collaborative character, individual efforts contributing to its general success. In science, as Bacon conceives it, truly effective results (not the illusory achievements of magicians and alchemists) can be attained only through collaboration among researchers, circulation of results, and clarity of language. Scientific understanding is not an individual undertaking" (Rossi, 1996, p. 32).

A rearrangement of Rossi's synopsis of Bacon's work would create a definition of COI:

> [A community of interest has a] public, democratic, and collaborative character, individual efforts contributing to its general success. ... [E]ffective results ... [being] attained ... through collaboration among ... [members], circulation of results, and clarity of language.

As the development of science progressed, Galileo Galilei (1564–1642) (Encyclopedia Britannica, 2006c) paid the penalty in 1633 for "circulation of results" to a community controlled by the Roman Inquisition (Hawking, 2002, p. 391). While Bacon's ideals were being put into practice, they were dangerous if not *properly* practiced!

Invisible Colleges: Scholarly Communities of Interest

Robert Boyle (1627–1691) (Encyclopedia Britannica, 2006b)—of Boyle's Law fame[3]—was the first person to publicly discuss the existence of an "invisible college" in letters to his tutor in 1646 (Lomas, 2004, p. 63). This invisible college was a group of philosophers including Boyle, which was headed by Dr. John Wilkins (1614–1672) (Royal Society of London, 2006b), first at Oxford University and later at Gresham College in London. These "philosophers" (scientists) were discussing, implementing and refining Bacon's proposed methods—particularly the

2 That is, the doctrine of propositions (Barnes, 1984, p. 28, 40), elements (fire, earth, air, and water) (Barnes, 1984, p. 1602), causality (Barnes, 1984, p. 333) and forms (Barnes, 1984, p. 1690–1694). Aristotle's ideas differed most sharply from modern ones in his belief that the universe had never had a beginning and would never end.

3 p*V=c or pressure times volume of an ideal gas at constant temperature is a constant.

experimental method—through participation in the inception of one of the world's most famous and long standing "COI," namely the Royal Society of London (Royal Society of London, 2006a). An idea of the profound effect Bacon had on the Royal Society comes from stanza five of Abraham Cowley's 1667 poem, "To the Royal Society" (Cowley, 1667):

V

From these and all long Errors of the way,
In which our wandring Predecessors went,
And like th'old *Hebrews* many years did stray
    In Desarts but of small extent,
Bacon, like *Moses*, led us forth at last,
    The barren Wilderness he past,
    Did on the very Border stand
    Of the blest promis'd Land,
And from the Mountains Top of his Exalted Wit,
    Saw it himself, and shew'd us it.
But Life did never to one Man allow
Time to Discover Worlds, and Conquer too;
Nor can so short a Line sufficient be
To fathome the vast depths of Natures Sea:
    The work he did we ought t' admire,
And were unjust if we should more require
From his few years, divided 'twixt th' Excess
Of low Affliction, and high Happiness:
For who on things remote can fix his sight,
That's always in a Triumph, or a Fight?

Chartered by King Charles II of England in 1662, the Royal Society has been continuously active in promoting, critiquing, and publishing scientific work until the present day (Royal Society of London, 2006a).

A lesson to be drawn from the history of the Royal Society is that scientific work includes a significant social component. "Community" is not something imposed from outside like management mandated teams, work groups, and quality circles, but is generated from within because social interaction is necessary for the process to work. Thus, invisible colleges did not disappear in the 1600s. Their persistence arises from the fact that collegial interaction is a fundamental part of the scientific method (De Mey, 1982).

De Mey (1982, p. 133) considers COI the "'natural' out-growth of the search for specialties" in scientific work. According to Lievrouw (1990, p. 66), "an invisible college is a set of informal communication relations among scholars or researchers who share a specific common interest or goal." Griffith and Mullins (1972, p. 959) observe that "communication and some degree of voluntary association are intrinsic in science" and they enumerate the following characteristics of invisible colleges:

1. Invisible colleges are associated with areas of fast scientific growth.
2. They are related to radical theoretical innovation or new methods.
3. The new ideas are developed according to well-defined procedures and within well-defined limits.
4. There are only a few geographical centers.
5. There is intensive interaction (personal face-to-face communication) among members.
6. There is an organizational leader.
7. There is an intellectual leader.
8. There is a high rate of turnover (3–6 year tenure).
9. The college has a limited lifetime (10–15 years).

The primary purpose of an invisible college is to foster communications among researchers and provide a rapid informal feedback mechanism to the development of scientific work.

Some (e.g., Finholt, 2002) have suggested that the Internet and virtual communities called collaboratories[4] can remove the geographical collocation requirement (item 4, above). This newer concept has not yet displaced the invisible college (Finholt, 2002).

As can be seen from items 6 and 7, above, invisible colleges have leaders and a structure. The structure may be elitist or revolutionary, according to Griffith and Mullins (1972), who cite Neils Bohr and the Copenhagen school of quantum mechanics as an example of an elitist college. They give B.F. Skinner and graduate students in psychology at Harvard University as examples of a revolutionary college (Griffith and Mullins, 1972, p. 961).

Formal Groups and Societies—the Royal Society of London (RS), the Geological Society of America (GSA), the Association for Computing Machinery (ACM), and the American Chemical Society (ACS)—are all examples of scientific societies that have been formally organized with specific membership requirements and formal publications. Many of them are discipline specific.

Communities of Interest Over Time

Informal COI become more formal with time, evolving into different organizational structures in step with the evolution of the scientific enterprise (Zuccala, 2006).

4 Wulf defined a collaboratory as a research "center without walls, in which researchers can perform their research without regard to physical location—interacting with colleagues, accessing instrumentation, sharing data and computational resources, and accessing information in digital libraries" (Finholt, 2002, p. 77).

This parallelism is in response to the evolving needs of a developing scientific paradigm. Table 1 shows the phases of development of a scientific activity. In table 1, each phase represents a different stage in the program's development. The first phase is a pre-paradigm phase where the most innovative and revolutionary results are obtained. Phases two, three, and four are subdivisions of the Kuhnian progression of a scientific paradigm (Kuhn, 1970). This progression is from inception, through the exercise of normal science, to a phase where the new things to be discovered are fewer (having been previously discovered), and new activity has diminished. This eventually leads to the maturity of the discipline. The vertical axis in table 1 represents the characteristics of a scientific activity observed in each phase, including institutional, sociological, methodological, and output constructs.

Table 1. Stages of scientific activity (after De Mey (1982, p. 150)).

	Phase 1	Phase 2	Phase 3	Phase 4
Cognitive content	**Pre-paradigm**	**Single Paradigm**		
		Normal science constructive applications	**Diminishing productivity**	**Maturity**
Methodological orientation	Originality, philosophical, pragmatic	Verification productivity non-philosophical	Consistency	Apologetic philosophical controversy
Literature	Innovative document(s) and preprints	Papers	Textbooks and domain specific journals	Journal Biographies
Social structure	None	Communities of Interest		
		(1) Invisible colleges	(2) Formal groups and societies	(3) Communities of Practice
Institutional forms	Informal	Small symposia	Congress and formal meetings	Institutionalization (university department)

In table 1, the boxes bounded by dashed vertical lines show when a particular activity *begins* in relative time, but not necessarily when it *ends*. The production of papers, for instance, begins in phase 2 but may well continue through phases 3 and 4.

In this table, special attention should be given to the social structure category. This category is represented over time by three specific constructs: invisible colleges, formal groups, and CoP. These constructs, in turn, are represented as subtypes or specializations of COI, with organizational structures and objectives appropriate to the phase of scientific activity in which they occur. CoP may be created prior to phase 4, although they generally represent a maturation of the discipline. This diagram differentiates the community of practice from the other community subtypes that have been developed by scientists from Bacon's era.

Compare the above view of scientific activity with a different one, as represented in figure1. The phases in figure 1 roughly compare to the phases of table 1. The vertical axis shows the different workflows rather than different characteristics. The figure shows the work that is occurring at a particular time. Within each phase, a set of iterations occurs with many activities going on at the same time. Each iteration within a phase represents the activity not as a single continuum but as a series of starts, stops, and re-tracings in the development of each workflow.

Collaboration occurs during each phase of the workflow. The specific workflow labeled "collaboration" in figure 1 represents the constructive criticism necessary to improve the

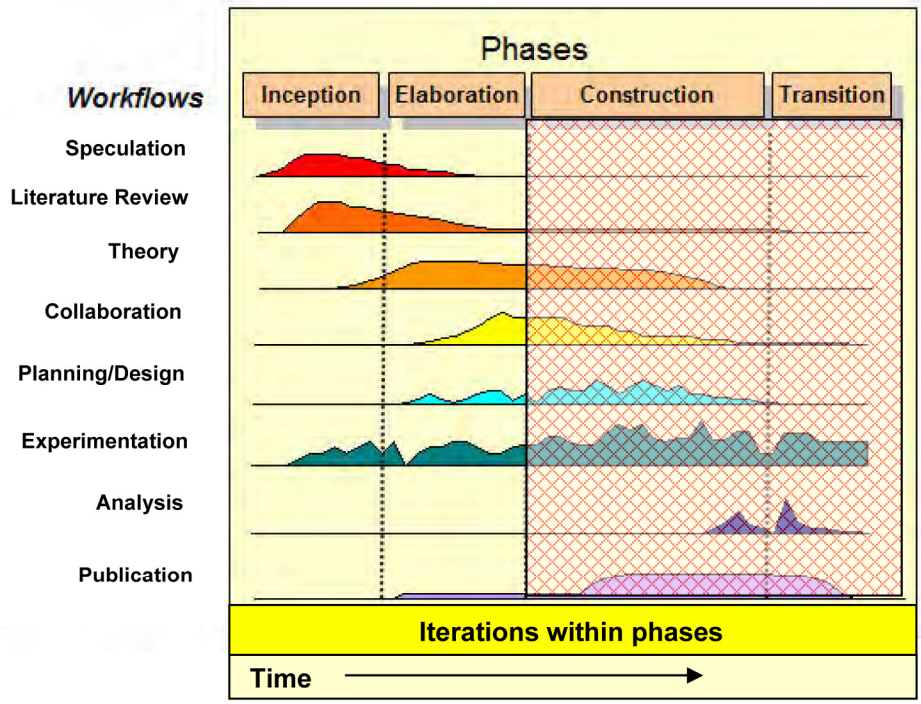


Figure 1. A scientific project plan, structured similarly to a software development project (after Wessberg, 2005).

theory so that an experimental plan can be designed. Speculation is done in concert with others, usually in small groups. Literature review is an indirect form of collaboration in which the researcher seeks out prior work that others may have done in the project domain. Theorizing and collaboration are performed together. The planning and design workflow usually includes a more formal "team effort" to allocate and marshal resources. Experimentation and analysis may specifically seek the criticism and collaboration of others. The publication workflow is designed to circulate the collective results to the community. In the model diagrammed in figure 1, CoP will form in the shaded area.

A scientific COI, according to Merton (1968), has not only a leadership but also distinct core and peripheral members. The dynamics of power between members of these two groups—the core and the periphery—follow what he called the "Matthew effect" (Merton, 1968, p. 58). The Matthew effect describes a situation where the more well-known members of the community tend to get recognition at the expense of the less well-known members. In addition, he proposed that "recognition" is the unit of exchange that is used within scientific communities to "buy and sell" intellectual property that is otherwise publicly available to all. In other words, intellectual property (e.g., ideas) is most freely exchanged if the exchange results in recognition for the discoverer. A scientist's intellectual wealth is acquired through the marketplace of COI (Merton, 1988).

Thomas Kuhn recognized that operational practice in the pursuit of scientific goals was part of the fabric of a community. He wrote:

> "A scientific community consists ... of the practitioners of a scientific specialty. To an extent unparalleled in most other fields, they've undergone similar educations and professional initiations; in the process they have absorbed the same technical literature and draw many of the same lessons from it. ...
>
> Communities in this sense exist, of course, at numerous levels. The most global is the community of all natural scientists. At an only slightly lower level the main scientific professional groups are communities: physicists, chemists, astronomers, zoologists, and the like" (Kuhn, 1970, p. 177).

This observation of Kuhn's suggests that CoP are a specialization of COI. As shown in table 1 and figure 1, CoP are created as a discipline matures—in other words, COI give rise to CoP because of specialization in the discipline and the need to share the accumulated knowledge at a practical, operational level.

Communities of Practice

The concept of CoP, like the concept of COI, is not new. The trade guilds of the Middle Ages could be considered COP (Wenger and others, 2002). Etienne Wenger explains the origin of the term:

> "The origin and primary use of the concept has been in learning theory. Anthropologist Jean Lave and I coined the term while studying apprenticeship as a learning model. People usually think of apprenticeship as a relationship between a student and a master, but studies of apprenticeship reveal a more complex set of social relationships through which learning takes place mostly with journeymen and more advanced apprentices. The term community of practice was coined to refer to the community that acts as a living curriculum for the apprentice" (Wenger, 2006).

In the view of Wenger and his collaborators:

> "Communities of practice are groups of people who share a concern, a set of problems, or a passion about a topic, and who deepen their knowledge and expertise in this area by interacting on an ongoing basis. ... Over time, they develop a unique perspective on their topic as well as a body of common knowledge, practices, and approaches. They also develop personal relationships and establish ways of interacting" (Wenger and others, 2002, p. 4).

Combining the idea of a CoP *as a learning activity* with the idea of a CoP *as more generalized collaboration* leads to the identification of CoP *as a specialized subset of* COI.

According to Wenger (2006), CoP have the following characteristics:

1. A common domain of interest.
2. A developed association of individuals that helps members engage in joint activities and discussions, help each other, and share information.
3. A practice consisting of a shared repertoire of resources: experiences, stories, tools, and ways of addressing recurring problems.

Some typical outputs that CoP achieve include (Wenger and others, 2002)

1. standards manuals;
2. improved skills;
3. reduced costs through faster access to information;
4. sense of trust; and
5. an increased ability to innovate.

CoP do not exist in a vacuum. The members of the communities also are usually part of another formal organization that provides them with the support to meet the basic components of Maslow's hierarchy of needs[5] (Maslow, 1970). Some of the relationships that communities can share with the formal supporting organization are listed in table 2. The most common relationship in scientific organization is the "bootlegged" type followed by the "legitimized" type.

5 From the most basic: (1) physiological needs; (2) safety needs; (3) needs of love, affection, and belongingness; (4) needs for esteem; and (5) needs for self-actualization.

Table 2. Relationships of communities to official organizations (modified from Wenger and others (2002).

Relationship	Characteristics	Typical challenges
Unrecognized	Invisible to the organization and sometimes even to members themselves	Difficult to see value and be aware of limitations, does not involve everyone who should participate
Bootlegged ... (The most well-known form of community of practice in science)	Only visible informally to a circle of people "in the know" A practical analog to the "invisible college"	Getting resources, sharing knowledge, training others, having an impact, keeping hidden, gaining legitimacy
Legitimized	Officially sanctioned as a valuable entity A practical analog to the "formal society"	Broader visibility, rapid growth, new demands and expectations, and pressure to perform in return for legitimacy
Supported	Provided with direct resources from the organization	Scrutiny; accountability for use of resources, effort, and time; short-term pressures
Institutionalized (by definition, ceases to be a community)	Given an official status and function in the organization	Fixed definition, over management, living beyond its usefulness

Communities of Practice Versus Teams and Other Management Structures

It is of crucial importance to remember that CoP are not task forces or teams. A task force is an organization created by management to perform a specific assignment. Once that assignment is completed, the task force is dissolved. A team, on the other hand, implements a specific process or function within an organization. A team structure addresses the interdependencies of different roles within that process or function. One of the best-known models of the team is owing to Tuckman (Smith, 2005).

Table 3. Comparison of group types (adapted from Nickols (2000), Hutchins (1985), and Wenger and others (2002)).

Group Types	Function	Basis of Membership	Basis of Cohesion	Duration
Formal departments	To deliver a product or service	Everyone or reports to the group's manager	Job requirements and common goals	Until the organization is reorganized
Operational work teams	Perform the ongoing work that has been assigned to the team	As assigned by management	Job/performance requirements and continuing, common goals	Until the work is completed or the organization is reorganized
Project teams and task forces	Accomplish a specific task or assignment, usually during a particular time frame	As assigned by management	Project milestones and goals	Until the project or task has been completed or the organization is reorganized
Quality circles	Identify, analyze, and solve problems in the workplace by interacting with management or solving the problems themselves where possible	Small group assigned by management	Product quality requirements	Ongoing, until the organization is reorganized
Communities of Interest	Collect and share information of common interest to be informed	Reciprocal value and acceptance, that is, members obtain and provide information of value. They may have specific membership requirements	Perceived value in belonging to a recognized group of peers and participating in a marketplace for the exchange of intellectual property	As long as members have a reason to interact, share information, and conduct constructive criticism
Communities of Practice	Develop members' expertise and define their place or role in the community	Self selected	Commitment and identification with the expertise that forms the basis of the practice	As long as members have an interest in improving the practice and maintaining the community
Informal networks	To receive and exchange information, to Know Who Is Whom	Acquaintances	Mutual need and relationships	As long as people keep in touch or remember each other

Please note: The grayed sections are controlled by management.

Table 3 lists seven types of groups, based on the function of that group in the organization, the basis for membership, what keeps the group together, and how long the group persists. The groups with a gray background are created and controlled by the management of the organization. The groups with a white background are not controlled by the organization and may be self-associating. An organization's management may support and facilitate the community groups but do not control them. The management-controlled groups are created based on the business needs of the organization. Communities are created because of the needs of the members of the community. Communities are of more value to the participants than the other group types. This value creates the incentive for them to be formed and persist.

Conclusion

COI and CoP fulfill different roles in the larger scientific community. Researchers in scientific communities of interest have, as a primary goal, creating and exchanging intellectual property for recognition within that community. Interactions among the members of invisible colleges and their more formal offspring, the scientific/technical/professional societies, will increase the value of the members' intellectual property in proportion to their recognition.

Figure 1. Starting time of community types relative to organizational maturity.

Members of CoP associate to engage in learning and skill improvement. The members of scientific COI also may be members of various CoP, as either "masters," "journeymen" or as "apprentices."

For members of CoP to be maximally supportive of the research community, their technical role in supporting the production and management of the intellectual property exchanged by researchers within formal societies (and elsewhere) should be recognized by all and not co-opted through enforced organizational mandate.

References

Bacon, F., 1605, The advancement of learning [Project Gutenberg online ed.], accessed March 2, 2006, at http://www.gutenberg.org/etext/5500

_____1620, The new organon, or true directions concerning the interpretation of nature, accessed February 17, 2006, at http://www.constitution.org/bacon/nov_org.htm

_____1626, The new Atlantis [Project Gutenberg online ed.], accessed March 2, 2006, at http://www.gutenberg.org/etext/2434

Barnes, J., ed., 1984, The complete works of Aristotle—The revised Oxford translation [2 v.]: Princeton, N.J., Princeton University Press, 2487 p.

Cowley, A., 1667, To the Royal Society, accessed March 12, 2005, at http://www.she-philosopher.com/library/cowley_ttrs.html

De Mey, M., 1982, The cognitive paradigm—An integrated understanding of scientific development: Chicago, University of Chicago Press, 316 p.

Department of Defense, 2004, Communities of interest in the net-centric DoD—Frequently asked questions (FAQ), accessed January 12, 2007, at http://colab.cim3.net/file/work/Expedition_Workshop/2005-12-06_Advancing_Information_Sharing_And_Data_Architecture/DoD/COI_FAQ.doc

Encyclopedia Britannica, 2006a, Bacon, Francis, Viscount Saint Alban, Baron of Verulam, accessed March 20, 2006, at http://www.britannica.com/eb/article-9108408

_____2006b, Boyle, Robert, accessed March 2, 2006, at http://www.britannica.com/eb/article-9016071

_____2006c, Galileo, accessed March 2, 2006, at http://www.britannica.com/eb/article-9105766

enTWIne Project, 2002, Community of interest [entry in online glossary], accessed February 21, 2006, at http://webguide.cs.colorado.edu:9080/entwine/Concepts/Concept19/viewDesc

Finholt, T.A., 2002, Collaboratories: Annual Review of Information Science and Technology, v. 36, p. 73–107.

Griffith, B.C., and Mullins, N.C., 1972, Coherent social groups in scientific change: Science, v. 177, p. 959–964.

Hawking, S.W., ed., 2002, On the shoulders of giants—The great works of physics and astronomy: Philadelphia, Running Press, 1264 p.

Hutchins, D., 1985, Quality circles handbook: New York, Nichols, 272 p.

Jubert, A., 1999, Developing an infrastructure for communities of practice—The Siemens experience, *in* McKenna, B., and others, eds., Online Information 99—Proceedings of the 23rd International Online Information Meeting: Oxford, Learned Information Europe, p. 165–168.

Kuhn, T.S., 1970, The structure of scientific revolutions (2d ed.): Chicago, University of Chicago Press, 210 p.

Lievrouw, L.A., 1990, Reconciling structure and process in the study of scholarly communication, *in* Borgman, C.L., ed., Scholarly communication and bibliometrics: Newbury Park, Calif., Sage Publications, p. 59–69.

Lomas, R., 2004, Freemasonry and the birth of modern science: Gloucester, Mass., Fair Winds Press, 386 p.

Maslow, A.H., 1970, Motivation and personality (2d ed.): New York, Harper and Row, 369 p.

Merton, R.K., 1968, The Matthew effect in science: Science, v. 159, p. 56–63.

_____1988, The Matthew effect in science, II—Cumulative advantage and the symbolism of intellectual property: ISIS, v. 79, p. 606–623.

Nickols, F., 2000, Communities of practice—Definition, indicators & identifying characteristics, accessed March 1, 2006, at http://home.att.net/~discon/KM/CoPCharacteristics.htm

Rossi, P., 1996, Bacon's idea of science, *in* Peltonen, M., ed., The Cambridge companion to Bacon: New York, Cambridge University Press, p. 25–46.

Royal Society of London, 2006a, Brief history of the Society, accessed March 2, 2006, at http://www.royalsoc.ac.uk/page.asp?id=2176

_____2006b, Wilkins; John (1614–1672) [biographical entry in online catalog], accessed March 15, 2006, at http://www.royalsociety.ac.uk/DServe/dserve.exe?dsqIni=Dserve.ini&dsqApp=Archive&dsqCmd=show.tcl&dsqDb=Persons&dsqPos=0&dsqSearch=(Surname='Wilkins')

Smith, M.K., 2005, Bruce W. Tuckman—Forming, storming, norming and performing in groups, accessed March 1, 2006, at http://www.infed.org/thinkers/tuckman.htm

Snyder, W., and Briggs, X., 2003, Communities of practice—A new tool for government managers, accessed March 1, 2006, at http://www.businessofgovernment.org/pdfs/Snyder_report.pdf

Wenger, E., 2006, Communities of practice—A brief introduction, accessed March 1, 2006, at http://www.ewenger.com/theory/index.htm

Wenger, E., McDermott, R., and Snyder, W., 2002, Cultivating communities of practice—A guide to managing knowledge: Boston, Harvard Business School Press, 284 p.

Wessberg, M., 2005, Introducing the IBM Rational Unified Process essentials by analogy, accessed March 3, 2006, at http://www-128.ibm.com/developerworks/rational/library/05/wessberg/.

Zuccala, A., 2006, Modeling the invisible college: Journal of the American Society for Information Science and Technology, v. 57, p. 152–168.

Knowledge Capture

Laure Wallace, U.S. Geological Survey

Knowledge capture is a critical step in turning data and information into reusable knowledge. While definitions differ, most agree that knowledge capture is an essential step in knowledge management—the process by which we create, identify, and distribute organizational knowledge to those who need it. It is not about capturing knowledge for knowledge's sake. The goals are (1) to increase the ability of people to share and codify best practices and information and (2) to create new knowledge. Power lies in the ability to share knowledge in a way that leads to continual learning in the organization and the synergistic creation of best practices for larger organizational benefit.

While technology can provide the means to organize and quickly access knowledge, critical knowledge also can be captured and shared in a variety of other ways. Mentoring, communities of practice (CoP), and joint problem solving between experts and novices are excellent techniques for sharing knowledge and the best practices on how to approach complex problems.

The U.S. Geological Survey (USGS) faces the loss of critical knowledge through retirements, attrition, and workforce trends toward multiple career jobs, unless we develop methodologies to preserve and enhance this knowledge. The dispersed nature of work argues for processes to share best practices and develop new understandings through groups such as CoP. The amount of time and resources available to experience and acquire personal knowledge is limited and makes the case for capturing and sharing understanding. Employees at all levels are in contact with many of the same customers, necessitating the need to have information about these interactions shared rapidly in order to leverage that knowledge to serve those customers better.

In rare moments, there are instances where people with great ideas and knowledge come together and are able to make something even better. We can no longer depend on the serendipity of those fortunate moments. Knowledge capture and management must be systematically pursued in the culture, shared, and used to help us adapt and grow, thus ensuring our continued scientific excellence.

With all of this in mind, a group met as part of the USGS Scientific Information Management (SIM) conference in March 2006, and tackled several critical questions to move forward the conversation on knowledge capture. The first question helped the group calibrate their understanding of the definition of knowledge. In the end, the group agreed that knowledge is more than information. It is a maturation process of digesting information within a given context, which includes the ability to apply a given understanding to some specific outcome. It is dynamic, valued, actionable, and includes a definition of the values, behaviors, and cultural requirements for successful action.

The drive to capture knowledge in the USGS is fueled by the prospect of losing knowledge through turnover and retirements, in addition to the potential inaccessibility of

information and data stored in everything from unlabeled boxes to obsolete electronic file formats. More important is the understanding that for our scientific mission to succeed, the USGS must be able to "re-use" the knowledge gained through past experiences, provide a mechanism to search valuable resources in a way that eliminates data overload, and understand the lessons that are learned by understanding knowledge evolution over time.

The USGS provides a powerful process for disseminating scientific information through traditional publications, websites, and CDs, but real knowledge about the business of science, the culture, the human experience, and understanding is most often shared through more ephemeral oral processes such as mentoring and storytelling, workshops, and conferences—and as such, may leave a gap in the knowledge chain for future reference and understanding.

The USGS faces several challenges in considering what knowledge to capture and how to capture that knowledge in a way that makes it easily accessible, relevant, and usable. Perhaps the biggest challenge is identifying which knowledge is most important to capture. Incentives to capture knowledge, beyond current scientific findings, are limited. Even in the realm of scientific information dissemination, many believe that a publication alone is the knowledge capture. Basic data and information on the processes and experience of developing that scientific publication are lost over time. To begin real knowledge identification, capture, and dissemination, the USGS will need to go through a culture change and provide the time and resources to ensure that we truly learn from the lessons of our past and thus aid in our future scientific growth.

If the USGS hopes to develop a robust culture for knowledge capture, it must address the human dimension of resistance to sharing knowledge, which derives from a reward system that gives credit for the *individual* development of data and information. Incentives must be created to reward the *collective* process of knowledge capture and sharing.

The USGS must begin to mine methodically the incredible data that already exists in a wide variety of mediums and to identify knowledge that has a continuing application to current scientific and business initiatives. COP must be diligently supported with the technological tools to capture and re-disseminate knowledge. A Chief Knowledge Officer should be established with a clear mission to develop the processes of identifying, collecting, and disseminating essential knowledge. Wiki technology should be explored as a mechanism for every employee to record critical organizational, scientific, and business knowledge. Finally, the USGS should explore opportunities to establish a process to capture knowledge from senior staff and staff in critical positions before they retire.

Knowledge capture—the process by which we create, identify, and distribute organizational knowledge to those who need it—is being recognized as one of the essential components for a flexible, efficient, effective, and robust scientific future for the USGS.

SODA—A Self-service Online Digital Archive

Rex Sanders, U.S. Geological Survey

Abstract

The U.S. Geological Survey (USGS) should implement a Self-service Online Digital Archive (SODA) to enable easy archiving of scientific data by scientists and technicians. This paper includes a description of the problem; a brief description of the proposed user-interface and data flow; and descriptions of benefits, drawbacks, features, and interesting issues.

Introduction

Point, Click, Saved Forever.

No, this isn't online salvation of your soul; it's online salvation of your data.

Problem

Lots of scientists and technicians have lots of digital data residing on floppies, CD-Rs, Jaz disks, internal and external hard drives, 8-mm tapes, and dozens of other formats. Nobody else knows about this data, and in most cases it's the only copy of the data. Data are frequently lost forever for a variety of reasons, including media failure, accidental deletion or disposal, and theft. Even if the data are not lost, much data are useless without the metadata, which is often locked in the faulty memories of the scientist or technician who collected the data. One scientist recently left the USGS, leaving behind hundreds of CD-Rs with minimal (if any) metadata about each disk.

Nearly 50 percent of USGS research scientists are eligible for retirement. Policies require archiving their work before they leave, but we have no infrastructure to support archiving digital data.

Sometimes our scientists and technicians have the best intentions: one scientist made three carefully labeled copies of each CD-R full of field photographs, for a total of 90 CD-Rs. Three years later, all of these CD-Rs were unreadable because the sticky labels had rotted the data layer.

USGS Western Coastal and Marine Geology (WCMG) has spent more than $500,000 in recent years to rescue data on rotting 9-track tapes. WCMG still has 20,000 tapes left to rescue.

As data and information managers, we need to make digital-data archiving as easy as possible for scientists and technicians. Ideally, we should make digital-data archiving easier than burning another CD-R.

Possible solution: SODA.

Not carbonated sugar water: a Self-service Online Digital Archive—SODA.

How will SODA work?

1. Point your web browser at the proposed website, *soda.usgs.gov*, and click the "Submit Data" button.
2. Select the data type and format to be uploaded.
3. Fill out a short form with the minimum metadata required (or more).
4. Select a simple release policy.
5. Click the "upload" button to send data to SODA through your web browser.
6. SODA will return a "guaranteed forever" URL, immediately accessible from inside USGS.
7. Later, an archivist will review your data, metadata, and release policy.
8. If everything is OK, the archivist will enable public access to the same URL (according to the release policy) and e-mail a notice to you.
9. If a problem is detected, the archivist will contact you for correction(s).

Benefits of SODA

- Increases internal and external access to our data.
- Enables scientists and technicians to archive data easily and immediately.
- Scientists can cite permanent URLs in publications and don't need to respond to data requests.
- Reduces the workload needed to archive data.
- Reduces the workload needed to rescue data from rotting media and eventually eliminates the need to rescue data.

Drawbacks of SODA

- Requires the scientist or technician to do something to archive data—information professionals won't be doing it for them. On the other hand, the information professional probably couldn't do it anyway, because (a) the metadata are usually locked in the scientist's mind, and (b) the resources aren't available.

- Requires one-time and continuing software-development costs to create a flexible, long-lasting tool that meets our needs. We are unlikely to find an off-the-shelf tool that could meet our needs without substantial customization or modifications.

- Requires ongoing support of petabytes of disk storage (including backups and (or) replication), offsetting some of the benefits of getting out of the "rescuing rotting media" game.

Other features of SODA

- SODA is not intended to replace existing, well-run data and information-management efforts in the various USGS disciplines and programs. SODA will catch all the "other" (a.k.a. "shoebox") data that has nowhere else to go.

- We must emphasize the user-centered design. If the users don't use SODA, we won't solve our problems. Using SODA must be as easy, or easier, than burning another CD-R.

- Disks are cheap, and backup and (or) replication is good insurance, so we will keep everything "online" forever. Variations will include automatic migration to nearline storage and replication in several centers to improve access and ensure disaster recovery—all depending, of course, upon funding.

- SODA: Our "Forever" Guarantee
  - If your data and metadata are reasonable, we will keep your data and metadata online forever.
  - "Forever" means we will take reasonable steps to ensure that your data and metadata are immune to human and natural disasters.
  - Reasonable steps include
    - Storing your data online in RAID 5 or better disk systems; this provides redundancy.
    - Backing up your data onto offline media.
    - Periodically sending backup media to secure offsite locations.
    - Keeping several generations of backups as security against malicious software and human error.
    - Using strong checksums to guard against accidental or malicious data modifications.

- Users can sign up for e-mail notices or RSS feeds when new data in their area of interest become available.

- The proposed website, *soda.usgs.gov*, could have two metadata search engines: one for internal-access-only data, the other for public-access data.

- SODA metadata should be based on XML for maximum reuse by other tools.

- SODA website design should be based on REST architectural style for maximum simple reuse by other tools—***not*** SOAP or so-called "Web Services."
  (For more information on REST, see http://rest.blueoxen.net/
  http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm)

- By simplifying the user interface and using a REST architecture, the implementation details can be hidden and changed as technology changes. It won't matter if the underlying structure is ASCII files with Fortran or Oracle with flocks of database administrators.

- Using REST, anyone can implement a number of services on top of SODA including

    - Multi-server search and retrieval: Even if USGS doesn't run one big SODA server, we could link discipline/program SODA servers and link to non-USGS SODA servers.
    - Map-based search and retrieval: Combine SODA metadata with tools like ArcGIS or Google Earth.
    - Mirrors: Automatically replicate data of interest from a variety of SODA servers to a local server.
    - Mashups: Directly re-use GIS coverages in ArcGIS; display images on maps using location metadata with Google Earth, integrate with National Map or Geospatial One-Stop, and so on.

- SODA must compute a robust digital signature (e.g., SHA-512) for each uploaded file and keep that checksum with the metadata to ensure that files are not accidentally or maliciously changed. SODA back-end software should periodically check digital signatures in the database.

  (For more information on the SHA-512 Secure Hash Standard, see http://csrc.nist.gov/publications/fips/fips180-2/fips180-2.pdf)

- We could enable easy internal desktop read-only access to SODA files through WebDAV, SMB, and other protocols.

- SODA must be designed to scale tremendously in several dimensions, including

    - Billions of files
    - File sizes above 1 terabyte (TB)
    - Total data storage size above 10 petabytes (PB), or 10,000 TB
    - Thousands of hits per second from automated tools

- SODA should be helpful to the user with hints, frequently asked questions (FAQs), and help buttons available throughout the website.

- SODA's user interfaces should be designed with several rounds of user testing.

- Users should be able to enter extended metadata (beyond the minimum required); users should be able to upload their own metadata files if they have them in several standard formats (e.g., FGDC standard, FGDC XML, SDTS, ArcInfo document.aml with conversion, and so on).

- We should never tell scientists they aren't allowed to upload files using proprietary formats (e.g., MS Excel); we should capture the files and metadata now and deal with these proprietary formats, rather than waiting for the scientists to retire and finding piles of CD-Rs in their offices.

- SODA should be "self-learning" as much as possible. If a user selects "Other" as a data type or format, SODA should do as much as possible to enable the user to describe and upload this data, while triggering archivist review.

- Archivists must be able to add new data types, data formats, metadata forms, and release policies, because our work changes over time.

- SODA's user interface and architecture must be designed as platform-neutral as possible, including using trailing-edge technologies, to ensure longevity.

- We should explicitly describe how reliable this archive will be so users will have confidence that using SODA is better than burning another CD-R.

- Maintaining online archives is much more expensive than hooking up another 500 GB hard drive, and this fact should be made clear so that users see the value added.

- SODA should be designed to last at least 50 years.

- Base philosophy: Online data are easier to keep around "forever" than offline, rotting media. When disk drives die you are forced to migrate to new disk drives, and you should be keeping good offline backups. New disk drives typically cost much less than old disk drives for the same capacity. For example, when your 1 TB RAID array dies in 4 years, you'll be able to buy a 16 TB RAID array for the same price and migrate with 94-percent expansion capacity. New developments in storage subsystems will lower the electricity costs of disk storage by putting banks of disks to sleep when unused (with periodic wakeup/integrity checking), just like putting laptop disks to sleep.

Interesting issues and further developments

- Versioning: What happens when the user wants to upload an "improved" version of a data set previously uploaded? Do we
  - Replace the earlier data set?
  - Keep the earlier data set but hide the metadata data from the search engine?
  - Add a "superseded by" link to the metadata for the earlier data set?
  - Keep both versions and let the users figure out which is newer?

- Multi-file data sets: How do we handle multi-file data sets? Some data sets are naturally multi-file (e.g., image data split into R, G, B files); some data sets are less clear (90 bathymetry files from one cruise). Do we require separate metadata forms for each file? Do we encourage zip, tar, and gz uploads with one metadata form?
- Rejects: How do we handle uploads that the archivist doesn't like—incomplete/incorrect metadata, inappropriate files, and so on?
- Release policies: We should face the problem we've been avoiding for years: standardized data-release policies. The USGS is working toward these policies, but the references aren't readily available. Ideally, we could distill release policies into a small set of menu choices: for example, 1, 2, or 3 years; internal or public release; with an option for the scientist/archivist to enable immediate release if the paper is published ahead of schedule.
- Metadata policy: Should metadata be publicly searchable immediately, with restrictions on data access until release policy says OK, or should metadata searches be restricted by the release policy as well?
- In general, we should try to work out as many policy issues as possible prior to development of SODA software.
- Preference for open-source software: Since the development work and benefits from SODA can be spread more widely using open-source tools and an open-source development—
    1. We should prefer a development environment based on widely available open-source tools—not necessarily Linux-based, but we should consider Apache, MySQL, Perl, PHP, and Python (the "AMP" in LAMP).
    2. We should open-source the software developed for this project, unless we want to play the Cooperative Research and Development Agreement (CRADA) or tech-transfer games.
    3. We should seek partners inside and outside USGS to share ongoing development costs and implementation experiences.
- A quick scan of other archiving tools shows that none are as simple as SODA; most are targeted at document management. We need a comprehensive survey of "what's out there," followed by a buy, modify, and (or) build decision.
- Ideally, GIO would implement SODA for the entire USGS.
- Rather than the USGS running the SODA servers, we should consider outsourcing server operations to a reliable company with good Internet access, a good server operations track record, good data-backup procedures, and options for replicated servers.
- Upload of data in certain common proprietary formats might trigger "automatic" translation to a preferred format. For example, Word, Excel, PowerPoint, Illustrator, or Photoshop files could be translated to PDF/A. Disk space is inexpensive, so we would keep both versions around until the proprietary format became unreadable.

III. Community Session Reports

During the SIM Workshop, 12 existing or potential communities were identified. Participants chose to participate in these various communities during two breakout sessions. The groups were

- Archiving of Scientific Data and Information
- Database Networks
- Digital Libraries
- Emerging Workforce
- Field Data for Small Research Projects
- Knowledge Capture
- Knowledge Organization Systems and Controlled Vocabularies
- Large Time Series Data Sets
- Metadata
- Portals and Frameworks
- Preservation of Physical Collections
- Scientific Data from Monitoring Programs

Comments and suggestions from these breakout groups have been captured below for the groups that met and contributed a session summary.

Archiving Scientific Data and Information

Notes captured by John Faundeen

Participants (36)

Joye Durant, Catherine Jarnevich, Susan Tewalt, Laurel Bybell, Anne Frondorf, Sally Cook, Stephen Snyder, Michele Banowetz, Keith Kirk, Phil Redman, E.J. McFaul, Rex Sanders, Jerry, Ornelas, Jana Stewart, Tom Cuffney, Carolyn Degnan, Dennis Mann, Carmelo Ferrigno, Rani Nandiwada, Nancy Soderberg, Steven M. Smith, Cheri Yoesting, Chris Polloni, Susan Ahrendts, Shawn Dadisman, Anthony McDonald, Marilyn Billone, Joe Langdon, Sharon Shin, Carolyn Reid, Jo Anne Stapleton, Kevin Foley, Clint Steele, Robert Matthias, and John Faundeen

Representation by Geographical Area

Reston, VA (16); Fort Collins, CO (2); Menlo Park, CA (6); Madison, WI (2); Santa Cruz, CA (1); Denver, CO (3); Raleigh, NC (1); Woods Hole, MA (2); St. Petersburg, FL (1); and Sioux Falls, ND (2)

Participants Unable to Attend Who Requested to be Part of the Community

James Hoffman (Denver, CO) and George Lienkaemper (Corvallis, OR)

Issues and Needs **(Captured by the Group)**

- Rescue data
- Archive for later use/accessibility
- Preserve context
- Providing a methodology for archiving
- Cataloging methods and tools
- Preserving the "working" data
- Documentation and metadata of the process and the data
- Archiving electronic records other than the ones National Archives and Records Administration (NARA) will accept [media/format]
- Assessment of records [appraisal]
- What constitutes an archive?
- Diversity of challenges
- Passion
- Solve a local problem
- We all have common issues and ideas to share plus records-management problems to solve.
- How do we preserve the data?
- What about visual data?
- Want to learn
- Best way to make it [data] available and preserve[d]
- Digital data
- Physical data
- Many pack-rats by choice or by default
- Effects of consolidations and management….not sure what to do
- Defining lifecycles of data types and technologies
- Cultural issues – raising awareness of our [needs to other] scientists and top of the organization
- Buy-in from management
- Obtaining agreements from NARA [on/concerning] our records
- Responsibility and division of labor…..is it transferred? To whom? Orphans?
- Enterprise database that tracks what was sent, when, etc.
- Identification, use and maintenance, and disposition [records lifecycle]
- Evaluation of data
- Quality and integrity of data
- Social security analog….[baby boomer analog?]
- Methods for archiving for scientists
- Process for submitting and maintaining in external [archives?] – (lifecycle)
- Can't pass on the problem to someone else….it is our responsibility

- Common understanding of what is happening within USGS
- Meeting NARA needs [and] our needs
- Administrative records not part of our concern

Issues **(Captured by the Small Group Breakouts)**

- Understanding roles and responsibilities related to archiving USGS scientific data and information – Rank 1
- How to deal with legacy data while addressing new and future data – Rank 1
- Media expires – Rank 1
- Usability of data depends upon metadata – Rank 2
- Resources are needed for data and information archiving
- Digital data volume is increasing with technology
- Who will preserve the data?
- Long-term access to digital/electronic [data] publications archival and preservation – Rank 1
- Collect/document project raw data, metadata, and background [information] to allow recreation – determine level of content – who stores the document and how it is retrieved – Rank 2
- Physical media/historical materials degradation issues
- Data file formats
- Archiving scientific data for dummies – Rank 1
- Metadata – archiving methods – Rank 1
- How many databases for metadata?
- Holistic archive
- Field record – linking physical, digital, [and] archival components
- USGS policy and guidance on records archiving – Records Schedule Disposition – Rank 1
- How to keep digital info on up-to-date media – Rank 2
- Ensure employees who are near retirement are preparing to pass on their materials
- What do you save? – Rank 1
- Ancillary data and methods
- Proprietary versus non-proprietary – Rank 1
- Archiving software – Rank 1
- Ownership – stewardship
- Convincing scientists to make their data available and usable – Rank 1
- Quantity of data – Rank 2
- Diversity of data – Rank 3
- Data of various formats
- Data on varying media
- Ensuring accessibility of data

Needs **(Captured by the Group)**

- Need to understand the scope of the challenge – Rank 1
- Sharing knowledge [of archiving scientific data and information] – Rank 1
- Need to have legacy requirements to read old media and migrate it to more current media
- Migrating data from media to maintain accessibility
- Capture and create metadata
- Incorporate data archiving in project and program plans
- Environmental requirements, backups, security resources, and funding
- Establish format standards
- Cross discipline format standards
- More NARA education
- [More on] NARA process
- [More on] NARA and USGS
- [More on] USGS archival [process, services]
- Resources including money and technological resources to stop materials decay (paper, other resources)
- Guidance – Rank 1
- Communication
- Where to develop?
- Index Catalog
- Living metadata
- Make sure employees know the USGS policy and guidance on records and archiving – Records Schedule Disposition – Rank 1
- Money to make [data] transitions
- Data interchange standards
- Responsible agents
- Decisions
- Methods
- Discipline-specific metadata
- Standard method for managing data from cradle to grave, including issues (quantity, formats, media, diversity)
- Better ways of integrating varying types of data sets
- Providing data in a readily readable format
- Money
- Centralized storage location for the data

Tools, Services, and Best Practices **(Captured by the Group)**

- Metadata creation tools and standards
- Standardized tool, best practices – collect highest resolution and then degrade to purpose
- Maintain mirror copies, utilize NARA

- Simple – SODA [Self-service Online Digital Archive]
- Custodianship
- USGS website
- Exit interview for retirees
- Shared facilities and standard procedures
- Better visualization tools
- Better ways of transferring data between media and formats
- Better ways of making data available
- Tools for ID-ing available data

Reported at SIM

Definition/Scope

Archiving scientific data and information is a broad topic—one that has big needs and involved issues. Consequently, the group struggled to define its scope and agreed to allow this to evolve.

Issues

The issues included the difficulty in dealing with legacy data with examples such as media obsolescence and the lack of access provided to researchers. Additional examples included the usability and context issues related to legacy data. Metadata is a big legacy-data area with the lack of, the sporadic nature of, the quality of, and the form that the legacy metadata takes affecting many of us.

Needs

Some needs identified by the group became glaringly apparent, including our USGS Records Management Program and how it relates to NARA. It was clear that more records-management training is needed.

Communication Vehicles the Group Will Use

E-mail	Chris Polloni and Joe Langdon, action to initiate
BBS	Desired
WebEx	Chris Polloni and Joe Langdon, action to initiate
Wiki	Jerry McFaul/GSA, action to initiate

Decisions Made by the Group

To establish a CoP
To be led by Chris Polloni and Joe Langdon

Database Networks

Session Facilitator, Chris Polloni; Session recorder, Rex Sanders

Prior to the workshop, 55 individuals expressed an interest in this community activity. We had 26 identified participants: 24 sitting at six tables, a facilitator, and a recorder.

Small group activities were organized by table designation as follows:

- Table 1—Scott McEwen, Sky Bristol, Kathy Lindblom, and Rani Nandiwada
- Table 2—Candy Bostwick, Dennis Mann, Tina Pruett, Kernel Ries, Roy Sonenshine, and Annie Simpson
- Table 3—James Sayer, Susan Tewalt, Ruth Jacobs, and Joan Freeney
- Table 4—Paul Geissler and Chris Rusanowski
- Table 5—Howard Kochman, Catherine Jarnevich, Annette Olson, Scott Wilson, and Peter Ruhl
- Table 6—Jolene Shirley, Shawn Dadisman, and Stephen Snyder

Workshop participants who indicated interest in our community and whose names came up often were Harry House and Jerry McFaul.

Our first group activity centered on capturing names of databases in use by the participants, and these included mySQL, Postgres, Ingress, Oracle, SQLserver, RDB, ACCESS/Excel and FilemakerPro.

Some general observations captured from the group

1. What's the one right answer?
2. Standards and architecture
3. Crosswalks, bridges, and interoperability
4. What's out there?
5. Time to look for standards
6. 20 years of frustration

Storage has grown tremendously – polled group for active database size

<100G - 7
<1,000G - 3
<10TB - 4
<100TB - 4
Over 100TB – 2

Issues and Needs (N for need, W for why, and I for issue) captured from the group

N – Want to interconnect databases
N – Easier to put stuff online
W – Don't need interop for interop sake
W – What are some of the science needs?

N – State-of-the-art techniques for data exchange – outside USGS
N – Step by step, how to setup database for interoperability – best practices
W – Need for interdisciplinary reports – synergy
N – Connections within projects legacy data
N – Need ecosystem view – across agencies
W – Don't know what's out there, don't want to duplicate effort
N – Need inventory of databases
W – Should be open to totally different uses of information
W – Change behavior of USGS/change culture/change reward system
N – Need incentives/disincentives:

- Withhold funding
- Backups
- Analytical tools

N – Must have metadata to have interoperability
N – Policy for open access to databases
N – Capacity planning
I – Networking databases – what's out there – Interconnect Standards
I – Never enough money to manage data within project – stolen for science
N – Inventory – what, what level, what services, lifetime, whom
I – Culture change
N – More meetings like this
I – Data should get a piece of the $$, like facilities
I – Inventory – who is responsible, planning capacity – Role of NBII/GIO Bureau level planning
I – How do we answer questions across databases to address larger issues
I – How to deal with archiving, contingency, real time, security – technical knowledge
I – Security
I – Provisional research data
N – Bandwidth in the field
I – Data quality – need to know
I – Data-use policies, e.g. right to use but cannot distribute
I – Taking risks – data quality
I – How to search many databases in different formats
I – Stewardship – homegrown, retirements
I – Update – migration to new technologies
N – Data-mining tools
N – Best practices to implement open-access policy
N – Money
N – Support
N – Inventory
N – Know what to serve the public
N – Inventories
N – Cataloging tools
N – Standards and archiving models for best practices
N – Define crosswalks
N – Metadata – what works

N – Tools – what's available, support
N – Harvest metadata
N – Analysis
N – Better defined protocols
N – Database of databases
N – Metadata tools – examples of what works
N – Success stories linking databases
N – Lessons learned
N – Incentives to researchers to relinquish data
N – Infrastructure
N – Management buy-in

Issues, Needs and Tools, Services, and Best Practices from each table with comments (N for need, W for why, and I for issue)

Table 1
I – Networking databases – distributed queries – linkage/relationship – standards inventory with data model
N – Policy for open public access to other databases – to avoid costly duplication – to dissolve barriers to merging data to form information
I – Overcoming security limitations
I – Capacity planning process
T – Data mining tools and techniques – to 'webify' data model visually
T – Access tools and techniques
BP – Database development and sharing best practices
N – Standards to enable policy compliance

Additional Comments from Table 1

- "Knowledge is the commonwealth of humanity."
- Why? Leverage the aggregate investment in data. To affect behavior from "hoard" to "share." To answer questions that cannot be answered today with disparate data sets.

Table 2
I – Funding, need top down approach – support – interoperability
I – Missing data in project area
I – Money allocated to manage data – control issues – re-use of data
N – Money – inventory – budget (central) – culture change
T – Web form to register data

Additional Comments from Table 2

- "Public databases should have open architecture."
- "Data providers cannot and should not predict how users will use it."

Table 3
I – Identify barriers to moving data to information
I – Getting past reservations to share information
I – Role of NBII and GIO – do they have complimentary or conflicting roles?
N – How do you get to the answers? N – Standards and architectures for best practice
N – Inventory, catalog tools of metadata for databases
T – MS-SQL 2000/2005

Additional Comments from Table 3

- "I would prefer to know and expose the metadata so I can decide if database has value."
- "Where is NBII and the GIO in helping USGS internally?"
- "Interoperability is needed at some scale."

Table 4
I – Why?
I – Purpose of this group?
N – How do we answer questions across databases/regions to address larger issues?
T – First step – inventory
T – Second step – cross-walk

Additional Comments from Table 4

- Qs Archiving?
  - Contingency? How to?
  - Bandwidth? Regs?
  - Real-time data? How to?
- Distributed databases? Data quality? Configuration?

Table 5
I – Security
I – Bandwidth
I – Data quality
I – Permission levels (data-use policies)
N – Access data in disparate databases
N – Know what exists (data discovery)
N – Harvest model
N – Methods, protocols in metadata

Additional Comments from Table 5

- "Desktop harvest model (Quick)"
- "Interested in integrating disparate data sets to facilitate distributed queries"

Table 6
I – What databases are out there?
I – Not re-inventing the wheel
I – Collecting metadata made easier
I – Not all databases need to be connected.
I – How to search various databases – many formats
I – Standards
I – Getting hands on data
I – Stewardship of databases
I – Updating databases
N – Database of databases
N – Better metadata tools
N – Examples of who has been able to network different databases
N – Get scientists to relinquish data.
N – Incentives to get data
N – Documentation of databases
N – Support

Additional Comments from Table 6

- "Geophysical/Geologic databases in ArcGIS – coverages, shapefiles, and grid formats/ASCII data of flightlines – gravity point data – cooperation with the National Oceanic and Atmospheric Administration (NOAA) and National Geographic Data Committee (NGDC)"
- "How to interconnect them – Do they need to be interconnected?"

At this point we queried the participants to see if we could come to a consensus for developing a leadership component. We found some willing individuals that would be part of a core group with a set of observers led by an initial core group leader.

Core Group coordinator – Roy Sonenshein
Core Group members

- Candy Bostwick
- Shawn Dadisman
- Paul Geissler
- Catherine Jarnevich
- Kathy Lindblom
- Annette Olson
- Peter Ruhl
- Annie Simpson
- Stephen Snyder
- Susan Tewalt
- Scott Wilson

The first round of ideas for the Core Group to tackle included

- What next?
    1. Sleep on it!
    2. Build an inventory – BASIS extract, e-mail, online form
    3. Describe our own databases
    4. Discuss culture change
    5. Compile success/failure stories

Possible sub-groups

- Data managers
- Interconnectivity
- MySQL
- Data dictionary

Next Steps – Core Group provides direction utilizing appropriate communication protocols, which may include e-mail, WebEx, and Wiki/TWiki solutions.

Digital Libraries

Notes captured by Steve Shivers (Facilitator)

Twenty-five interested participants gathered at the SIM Workshop to discuss the formation of a Digital Libraries CoP within the USGS. The group included representatives from the Geospatial Information Office, the Geology Discipline, and the Biological Resources Discipline. A number of participants were formerly assigned to the Geography Discipline, so only the Water Resources Discipline's interests were absent from the discussion.

Participants brought a wide range of expertise and geographic diversity to the conversation; all regions were well represented. The group included librarians, information technology specialists, managers, and research scientists. Everyone seemed comfortable in the open forum environment. The discussion was cordial and occasionally passionate.

Prior to the workshop, those who indicated an interest in the digital libraries topics were provided with the following introduction:

> Scientific research requires high quality library support to succeed. Library science is undergoing a digital revolution that will have an effect on everyone in the USGS, yet the digital library as it relates to USGS is not clearly defined. To some it may refer only to the digitized holdings from a traditional library collection. In reality it is probably much more than that. Any collection of digital information along with the tools and support staff that make that information useful to its users could be considered a digital library. The types of information included in a digital library are only limited by the imagination.

> Cost is the overriding factor in moving from traditional libraries to the digital world. This community can search for common ground in an effort to develop digital libraries across the USGS. Developing a consensus on what constitutes a digital library within the USGS would benefit all parties involved, as would bringing together library professionals and those who have digital library needs in their programs. Gathering common requirements and developing some uniform strategies to meet those needs might be an achievable goal. This community may develop into a united pool of knowledge and current resources within the Bureau that could lead to appreciable cost savings as developments continue.

Much of the discussion during the session centered on what the term digital library meant to the individual participants. Each person was asked to provide one statement that captured a concept that was important to them. These comments were captured on flip charts to be grouped and analyzed at a later date. A great deal of subject overlap was recognized with the metadata and knowledge-organization-systems communities.

While the group did not have time to reach a consensus on a definition of digital library, the following key concepts were affirmed by all:

- Digital libraries should be geospatially enabled wherever possible.
- User feedback is necessary to ensure proper implementation. Developers need to define "users" and determine their requirements.
- Access issues must be addressed and resolved.
- Standards must be established, used, and enforced to ensure interoperability.
- Digital libraries include both traditional and non-traditional collections and services.

The group would like to continue with the formation of a CoP. Steve Shivers will act as coordinator until the National Library Coordinator position is filled. That person may become the champion for development of digital libraries in the USGS.

E-mail will be used for community communication. A group e-mail address has been established, which contains all session participants and those who indicated interest but did not attend. Participants can send messages to the list and indicate in the subject line specific information such as FYI, Input, or Action. Other communication tools that may be of use to the community will be investigated and tested.

An active Digital Libraries CoP will provide a strong unifying affect on developments within the USGS as we work to bring our information resources to their users' desktops. Working together provides our best chance for success.

Field Data for Small Research Projects

Session Facilitator: Fran Lightsom

Group members: George Lienkaemper, Janet Cushing, Martha Garcia, and Carol Simmons

1. Topic scope: What is a "small" research project?
We decided that the essential characteristic of a "small" research project is that it collects data for the purpose of answering a scientific question in the near term, instead of collecting data to create a database that could be used for multiple scientific purposes later. This often goes along with low budgets for staff and equipment, short duration, and a tight focus of attention. We suspect that the majority of projects in the Biological Resources Discipline and many in the Geology Discipline fall into this category.

We are involved in multiple activities that manage the scientific data from small research projects, including creation of metadata, preservation of the data, and making it available through data catalogs or indexes. These activities are important because the data have value beyond the purpose for which they are collected, and because the scientific method requires preservation of data for independent verification of interpretations and results.

2. Identification of best practices was beyond the scope of this meeting; however, we concluded that much of the data from such projects is probably undocumented (by metadata) and at risk.

3. Tools needed for information management.
(a) The Geologic Discipline database of databases is a valuable tool that should be replicated for the other disciplines. This would be a good task for the GIO to undertake.
(b) SODA (a Self-service Online Digital Archive), as proposed by Rex Sanders, would be a useful alternative to current small scale, high risk, data-storage strategies.

4. Value of a potential CoP.
During the meeting, a spontaneous discussion about gathering metadata demonstrated the value of exchanging information about successful strategies within a CoP. We hope to develop Internet systems for carrying on discussions and developing recommendations for various aspects of managing data for and within small research projects; a blog and a Wiki were suggested. It will be important for these to be available outside the USGS Intranet.

A valuable first project would be compilation of a list of currently (2006) available metadata standards, tools, and indexes appropriate for different data types. These may be found outside or within the USGS.

The CoP should include information specialists, research scientists who work on small projects, cooperators (including students), and retired professionals in these categories.

5. Expectations of GIO assistance.

- Internet-communication services (Wiki, blog)
- Further face-to-face meetings like this workshop
- Publicizing the existence of this COP and other CoP on related topics. Two opportunities for spreading the word are presentations at the Biology Managers' Meeting in August 2006 and the Global Change researcher's meeting in June 2006.

6. Participation.
Fran Lightsom agreed to be the coordinator, and Carol Simmons agreed to be in the Core Group. George Lienkaemper and Janet Cushing are interested in belonging to the CoP, and Martha Garcia will be our interface with the Office of Science and Technology Policy (OSTP) working group on scientific collections.

7. Recommended next steps.

- We will spread the word among our contacts and recruit additional members.
- We will connect with other breakout groups that are on topics of interest, especially metadata, archives, and digital libraries.
- We will propose working with the metadata CoP to begin compiling an inventory of metadata standards, tools, and indexes appropriate for different data types.

Knowledge Organization Systems and Controlled Vocabularies

Summary of workshop breakout session, captured by Peter Schweitzer

A few more than 20 people attended the session.

The topic overlaps considerably with both Metadata and Digital Libraries. Knowledge Organizational Systems and Controlled Vocabularies (KOS-CV) can be regarded as infrastructure supporting these communities because those activities tend to use KOSs, sometimes implicitly as well as explicitly.

The CoP includes developers of vocabularies and ontologies, catalogers, web architects, metadata writers and readers, and web users. Those in a nurturing role tend to be developers of vocabulary or software that manipulate the vocabulary and developers of applications (including website architectures) that make effective use of KOSs.

Members of the CoP chiefly seek information about how the activities of others relate to their own, specifically how to make use of KOSs developed by others and, to a lesser extent, how to contribute to the development of KOSs that are already ongoing. We look for ways to enhance the interoperability of our technologies as well as understanding the various KOSs available and what they are intended to do.

The CoP will initially work through e-mail. While we are not opposed to web-collaboration tools, it is not yet clear that they are what this CoP wants.

Metadata

Notes captured by Sharon Shin

A lively group of 20 convened to discuss metadata issues, best practices, needs, and next steps. The group decided to move forward with a CoP with Sharon Shin, Federal Geographic Data Committee (FGDC), and George Lienkaemper, Forest and Rangeland Ecosystem Science Center (FRESC), leading for a limited time period.

Issues:
Cultural barriers impede metadata as a component of data management.

- Managers do not value metadata.
  - Unfunded
  - Overlooked
  - Not created without support
- Metadata policy enforcement
- Metadata is not built into Center culture
- Metadata rejoined with data as a complete package
- View metadata as collaboration tool
- Metadata for legacy data sets
  - Level of detail
  - Fiscal support
- Appreciation and understanding of metadata
  - Public
  - New employees
  - Research scientists
  - Data partners

Tools

- Tool access/development that allows integrated metadata creation throughout project lifecycle
- User-friendly creation tools
- User-friendly discovery tools
- Metadata tools aside from GIS
- Dictionary builder–unique terms discovered in metadata are cued for addition to the dictionary builder

Standards

- Enforcement
- Apply the correct standard for the product or data type/set

- Establish content recommendation
- Implementation

Education/Outreach

- What has the USGS done?
- Assist in implementation and culture adoption
- Create a "How to Guide" or a "User-Friendly Metadata Guide"
- Share success stories from Centers with established metadata programs

Best Practices—The seasoned metadata community shared the following:

- Don't make scientists learn the standard
- Include metadata in project planning
- Gain management's metadata buy-in
- Metadata into Center Culture – field-sheet information migration to metadata
- Alternative metadata presentation – FAQ style to avoid learning the standard
- Use taxonomy/ontology to
    - ease metadata creation
    - avoid misspelling
    - improve discovery

Needs

- Metadata policy enforcement
- Follow through on metadata requirement prior to publishing
- User-friendly metadata – better tools
- Metadata in practice (for managers/scientists)
    - how, in time of reduced budgets
    - document data efficiencies (Office of Management and Budget (OMB))
- Utilize Information Technology Specialists
    - make metadata creation easier
    - update/implement new standards
- Tool review – what exists, ease of use
- Accountability in credit – make metadata count for research scientists
- Metadata outreach materials

Preservation of Physical Collections

Breakout Session Facilitators: Martha Garcia and Jerry McFaul

Breakout session attendees: Marilyn Billone, Laurel Bybel, Mike Frame, John Faundeen, Ron Lofton, Peter Lyttle, Brand Niemann (U.S. Environmental Protection Agency (USEPA)), Clint Steele, and Nancy Soderberg

General Observations

- It was good to have a session focused on physical collections as more attention focuses on digital-data collections.
- A "physical collection" needs to be defined. There are a couple of efforts underway by policy of the Office of Science and Technology and as mandated by the Energy Policy Act of 2005 that will help to formalize physical collection efforts. The USGS needs to have discussion beyond those activities to further define what physical collections need to be maintained.
- There are a number of tools that have been developed to suit specific needs that should be evaluated to facilitate uniformity among Bureau-wide collection activities. An inventory of all USGS collections needs to be done.
- There have been similar data-gathering activities for historic photos, books, and manuscripts. USGS employees involved with those activities should be interviewed to discuss lessons learned. Potential contacts are Nancy Blair, Bob Bier, and Greg Allord.
- The USGS has collection policies. In the past, it seemed that retiring employees did a better job of leaving useable collections. Those policies need to be reiterated and enforced. The group recognized that we don't have the resources to maintain all collections. Setting enforceable standards will help to make some tough decisions regarding the retention of collections. Warehouse collections should be reviewed.
- Small data collections need to be evaluated. Should they be merged with large collections? Collections should be maintained by people that have an interest in the collection.
- Business plans are needed for all collections that include staffing recommendations and potential partnership opportunities.
- Guidance is needed on when collections become museum artifacts, and appropriate steps should be taken to preserve the artifacts. Potential contact is Susan Russell Robinson.
- A Wiki is needed to link the community. Additionally, a Wiki could be used as a purge alert to get the word out on collections that are in danger of being lost as employees retire or leave the USGS.
- The CoP should take advantage of technological advances and use facsimiles as appropriate to replace physical collections.

CoP Coordinators: The Geologic Discipline is in the process of hiring a full-time program coordinator for data preservation. Peter Lyttle, Jerry McFaul, and Martha Garcia agreed to act as temporary coordinators until that position is filled.

GIO's role: GIO can provide uniformity and standards by assisting science and helping to manage information. Following the SIM workshop, the GIO should focus on a few CoP. This group has a good chance of succeeding as a full-time person will be dedicated to the effort.

Scientific Data from Monitoring Programs: The USGS Monitoring Community of Practice

The USGS Monitoring CoP (http://biology.usgs.gov/status_trends/MonitoringCommunity/) seeks to improve USGS monitoring through collective learning. We are currently on hold for the 2006 field season, but plan to resume activities in the fall.

We have

- 272 participants on our e-mail list
- conducted two surveys of participants to identify issues
- formed a core group, representing all disciplines
- established a website and listserver
- held two conference calls/WebEx on monitoring issues

Please see the website for more information.

IV. Panel Discussions

Scientific Information Management Workshop—Panel I

Perspectives on the Workshop: Impressions, Issues, and Implications

Remarks by John Faundeen

Crosscutting Issues

- Retirees are taking knowledge, and sometimes data, with them.
- Legacy data problems are large and pervasive.
    - Frustration over inability to address
    - Part of our scientific legacy is fading away.
    - Data rescue efforts are needed before it is too late.
    - Some attendees suggested inventorying the data at risk as a first step.
- Metadata challenges persist.
    - Not just with legacy data
    - The context of the records becomes lost without good metadata.
    - We must convince scientists and managers of why and how to create good metadata in order to be accountable.
    - We must institutionalize this as part of the project planning process and make it part of our culture.
- Sponsorship of data collections
    - When programs or projects change or end the stewardship of the data is often lost.
    - Who becomes responsible? This is not just a money issue.
- We need to better understand the USGS Records Management Program and the relationship with NARA.
    - The lifecycle of records still is not understood.
    - NARA's role is confusing to many.
    - Perhaps, we need a USGS Records Management "Geek Squad," which is prepared to address our scientific records-management challenges.

Emergent Issues

- There is no long-term plan for preserving physical specimens.
- The volume of electronic records is becoming more pervasive across our units.
- Efficient access to all of our data is still a concern.
- Obsolescence of media is affecting most staff.
- From the registration comments we noted around 40 percent indicated the need for better data management and preservation of our data. An additional 30 percent relayed that access to our data is still an issue, both for us and for others.
    - Perhaps we need a USGS Preservation Program.

Practices

- We need a process to review our records.
- There is a desire to utilize records-management practices.

Sharing Tools

- Conducting trade studies was discussed as a valuable activity with the results being shared.
- Outside USGS contacts also were discussed. USEPA was very involved in our workshop and the National Institute of Standards and Technology (NIST) may have resources that can be exploited.
- We shared some current (2006) online applications with the attendees that perhaps can be used in other areas of the USGS.

Summary

Lots of challenges exist and many of them are not new. What I was NOT surprised to find throughout the 3 days of the workshop was the enthusiasm for our science collections and the passion to preserve our scientific legacy.

Scientific Information Management Workshop—Panel II

Reflections on this Workshop, Communities of Practice, and a Culture of Resilience

Remarks by Fran Lightsom

On September 19, 2005, the *Boston Globe* published an opinion piece called "Fixing Government after Katrina." The author was Yossi Sheffi, a professor of engineering at the Massachusetts Institute of Technology, who has studied what makes businesses and other organizations successful in dealing with crises—why some succeed and others fail when unexpected events happen. Sheffi begins, "The response of government to Hurricane Katrina is being dissected to determine why the initial reaction was lackadaisical even though officials knew the disaster was coming. One reason could be the culture of the organizations involved." Sheffi goes on to list three cultural characteristics of what he calls *resilient organizations*.

- "Empowerment of front-line employees. … Front-line employees are close to the action and can assess what is needed; as a disruption develops there is usually not enough time to go through the usual chain of command." Empowerment means having both the authority and the ability to do what is needed.

- "Constant communications. Resilient enterprises communicate obsessively and ensure that they can communicate in a disaster. … Resilient organizations not only have the gear; they create the environment in which communications are important and bad news travels fast." Sheffi's last remark is worth emphasizing: bad news *should* travel fast.

- "The big picture. Employees in resilient enterprises are passionate about their mission and care deeply about what they do."

I kept a copy of Sheffi's article because his *culture of resilience* struck me as a good description of our culture for success in the USGS Coastal and Marine Geology Program. Our program's culture is based upon the experience of doing scientific research on a ship at sea. On a research ship, success requires self-sufficiency. When something breaks or fails, or the weather goes bad, you have to deal with the situation using the things and people (whatever their skills) you thought to bring with you—and things do break and fail. Sometimes, so many complications and problems crop up that the scientific crew simply would not bother *if they were not very passionate about doing science*.

Today I realize that this workshop, and the CoP it has fostered, can lead to a *culture of resilience* for SIM at the USGS. I will, therefore, structure my remarks around Sheffi's three points.

First, *empowerment of front-line employees*. We at the workshop are the people who do the work of SIM: the front-line employees for this job. The communities we are talking about are CoP—groups of people who are engaged in particular kinds of work and who get together to learn from and with each other. CoP empower front-line employees by developing their ability to do their jobs in new and better ways. Besides ability, empowerment also requires authority. Perhaps I should not mention it here, but CoP can be a challenge to the existing organizational culture. If our managers let us nurture these communities, improve our abilities, and learn new and better ways of doing our jobs, they will be implicitly giving us authority to work smarter instead of just harder. We will be empowered. SIM is done in every science center, in every team, in every project in the USGS. Many of these front-line employees could not travel to this workshop. As we consider next steps, we must find ways to include them.

Sheffi's second point is *constant communications*. It is interesting that Sheffi emphasizes communicating the *bad* news. During the workshop we heard about a New England invasive species network, which discovers the intruders early and spreads the news while the population is small enough that the invasion can be stopped. In the invasive species network, "bad news travels fast." In SIM, our "invasive species" are things like CD sticky labels that kill data or—to cite a more positive example—an opportunity to challenge standard operating procedure by incorporating new technology (for example, using Google Earth to showcase our information). In our CoP, we can talk about issues that challenge us. Unlike supervisory/accountability structures, with their pressure to report efficiency, effectiveness, and good news, a CoP gathers for the purpose of discussing interesting problems and better ways of solving them. When a troublesome problem comes to the attention of a USGS-wide CoP, the bad news travels fast. At this

workshop and in CoP, communication is what we do, but we have only begun. As we consider next steps, we must continue communicating if this workshop is to be a success.

Sheffi's third point is *the big picture and passion for the mission*. Being at a workshop, with all sectors of USGS represented, is seeing a big picture. When I make new friends who are biologists and realize that they are facing some of the same information-management challenges that I face as a geologist, I see that my everyday experiences are part of a larger mission. When Acting Director Leahy spoke to us on Tuesday morning, he showed us the big picture. As I return to my job, taking time out of my routine tasks to participate in a CoP will remind me of the big picture. Talking to community members from other disciplines and regions, who also care about SIM, will remind me of my passion for our mission. Look at our workshop goals: long lists of tools, techniques, and best practices. How easily we get busy with the details! If we are to succeed as a resilient enterprise, we need to collect the stories and images that remind us of our overall mission. As we consider next steps, should we have a "community of communities" to continually hold up the big picture of SIM in case we lose focus?

In summary, CoP for SIM will help us make the USGS a more resilient organization by empowering our front-line employees, improving communication, and keeping us all enthusiastic about the USGS mission. To get these benefits, we need to keep communicating among ourselves, we need to extend our lines of communication to include those who could not travel to this workshop, and we need a "community of communities" to maintain our focus on the big picture of USGS SIM.

V. Contributed Abstracts

1. “National Water-Quality Assessment Program Invertebrate and Algal Data Analysis Software” by Thomas F. Cuffney
2. “The Geodatabase Solution to Data Management: Examples from LASED and XSTORMS” by Shawn Dadisman, Karynna Calderon, Robert Wertz, James Flocks, and Janice Subino
3. “Using Web Metrics at the U.S. Geological Survey” by Kit Fuller
4. “Geospatial One-Stop Community Geographic Information System Portal Application” by Steven Hale
5. “The Marine Realms Information Bank—A Coastal and Marine Digital Library” by Frances L. Lightsom, and Alan O. Allwardt
6. “U.S. Geological Survey Science Topics Index and Supporting Infrastructure” by Peter N. Schweitzer
7. “Federal Information and Investment Programs” by Carmelo Ferrigno, Deborah Kimball, Amy Berger, and Judy Snoich.
8. “U.S. Geological Survey Enterprise Architecture—Utilization of the Data Reference Model to Improve Data Sharing” by Raymond C. Obuch and Stuart Doescher
9. “A Global Organism Detection and Monitoring System” by Catherine S. Jarnevich, Thomas J. Stohlgren, James J. Graham, and Gregory J. Newman
10. “Development and Utilization of a National Geochemical Database” by Steven M. Smith and David B. Smith
11. “The South Florida Information Access (SOFIA) System” by Heather S. Henkel
12. “myUSGS Portal Pilot Project” by Sky Bristol
13. “Geographic Information System for the Gulf—ADS40 Imagery on Lidar for Hurricane Katrina” by David Greenlee
14. “StreamStats: A Web Application for Streamflow Statistics and Basin Characteristics” by Kernell Ries
15. “Collaboration in the Development of the Bird Banding Laboratory Database System” by Kevin Laurent

National Water-Quality Assessment Program Invertebrate and Algal Data Analysis Software

Thomas F. Cuffney[1]

[1]U.S. Geological Survey North Carolina Water Science Center, Raleigh, N.C.

The U.S. Geological Survey (USGS) National Water-Quality Assessment (NAWQA) Program is a long-term monitoring program that integrates physical, chemical, and biological data to assess water-quality conditions across the conterminous U.S. Over the past 16 years, the NAWQA Program has sampled more than 15,000 sites and measured more than 2,800 parameters that describe land-use, land-cover, hydrology, geomorphology, water chemistry, and biology. Providing access to these data for USGS scientists and the public has been a challenging process. Initially, the NQWQA Program lacked a national database for storing biological data so data were stored at the project level and periodically aggregated nationally. This made it difficult to monitor data entry, data quality, develop data analysis tools, and meet deadlines for regional and national reports. The Biological Transactional Database (Bio-TDB) was developed in 1999 as a national repository for NAWQA biological data. Bio-TDB provides the ability to monitor data entry, evaluate data quality, and export data in a consistent data structure that is utilized by data analysis tools for invertebrates (IDAS), algae (ADAS), and habitat (HDAS). IDAS, ADAS, and HDAS provide nationally consistent and documented tools for inspecting, editing, and analyzing large data sets. The development of the NAWQA data warehouse (DW) in 2001 provided a centralized repository for all NAWQA data. This has greatly facilitated the acquisition of regional and national data. The combination of Bio-TDB, NAWQA DW, and data analysis tools allow NAWQA biologists to do in a matter of hours or days analyses that used to take weeks or months. Data analysis tools that were developed as stand alone programs (IDAS, ADAS, and GRAN) for NAWQA also are being utilized by scientists within and outside the USGS for the analysis of non-NAWQA data. In contrast, data analysis tools that were developed as an integral part of Bio-TDB (HDAS) are limited to NAWQA data stored in Bio-TDB.

The Geodatabase Solution to Data Management: Examples from LASED and XSTORMS

Shawn Dadisman[1], Karynna Calderon[1], Robert Wertz[1], James Flocks[1], and Janice Subino[1]

[1]U.S. Geological Survey Florida Integrated Science Center, St. Petersburg, Fla.

The U.S. Geological Survey (USGS), Center for Coastal and Watershed Studies, has developed a multiple-geodatabase system to manage decades of digital and analog data collected from the coastal zone. Presented here are two examples of project data that are managed by the geodatabase system: LASED (Louisiana Sediment and Environmental Database) and XSTORMS (coastal oblique aerial photography and videos of eXtreme STORMS).

LASED is the result of combined efforts of the USGS and academic collaborators to manage geologic data from the Louisiana coastal zone. The database incorporates a wide range of data types (sediment-sample logs and analyses, geophysical profiles, raster-image base maps, logbooks, etc.) that are integrated with spatial data to provide processing and visualization capabilities using standard Geographic Information System (GIS) and Internet-browsing tools.

XSTORMS is a recent exercise to rescue analog oblique aerial photographs and videos of the coast collected before and after major hurricanes. These data are spatially linked so that pre- and post-storm comparisons can be quickly made and the results shared electronically.

Benefits to storing project data in a geodatabase are numerous. They include centralized data storage to serve as a multi-user online data archive, routine backups and consolidated offsite storage, integration of different data types, and a project resource and analysis tool. Full access to the geodatabase system is available to registered users through the USGS Intranet, and limited access will soon be available to LASED through the Internet.

Using Web Metrics at the U.S. Geological Survey

Kit Fuller[1]

[1]U.S. Geological Survey, Reston, Va.

Data from several sources tell us a lot about the U.S. Geological Survey (USGS) Web presence, including visitation numbers (traffic), customer types, customer interests, satisfaction scores, and information about individual pages. Data are being gathered from USGS Web logs, USGS Search logs, USGS Frequently Asked Questions, the American Customer Satisfaction Index satisfaction surveys on USGS Websites, the Nielsen//NetRatings visitation and demographics database, the AccMonitor Section 508 checker, and the USGS Web Inventory and Registration System. Data about selected USGS Websites also are available from other sources, including the Maxamine Website crawler and recent usability studies of parts of the USGS homepage suite.

Each source of information has a different purpose, methodology, and value. Some sources of information generate a set of data that can be difficult to compare to other data. Some data sets include information USGS can act on immediately, and other data sets provide a general overview that can be useful in many ways. Also, different methodologies "see" different segments of the USGS Webscape. Used together, the suite of information sources presents a good general overview of the USGS Webscape, as well as specific "actionable" details.

EWeb has begun a Web Metrics project, which is taking a bureau-wide approach to Web metrics. EWeb has bought access to Nielsen//NetRatings data, which presents a high-level view of government and business Website traffic. Visitor demographics data from Nielsen characterize the kinds of general public visitors to USGS Websites. EWeb also has bought into the American Customer Satisfaction Index, which uses Web surveys presented to about 1 percent of the visitors to the USGS homepage area (although EWeb plans to expand the survey to operate on as many USGS Websites as possible). NatWeb provides the base of Web log data, augmented by logs from other major USGS Websites. Data are compiled for summarization and future uses and are analyzed by the Web Metrics Working Group (WMWG), a diverse group of about 12 USGS employees that represent all disciplines and regions and include technical gurus, content managers, and customer service experts. The WMWG compiles data on a continuing basis and updates the summary-data compilations, which are accessible on the Intranet. The WMWG also compiles a quarterly USGS Web by the Numbers report to summarize USGS Web activities and trends.

A Web Metrics project website provides access to the data collected and the reports and presentations of the project, as well as contact information and information about how the USGS can use Web metrics data to improve the effectiveness of USGS Websites.

Geospatial One-Stop Community Geographic Information System Portal Application

Steven Hale[1]

[1]U.S. Geological Survey, Reston, Va.

The communities of practice within Geospatial One Stop (GOS) have asked for and will require Geographic Information System (GIS) collaborative tools to carry-out analysis of GOS-cataloged geospatial data. The former GEODE team has begun developing advanced user-friendly enterprise applications that will address this requirement. The use of these geospatial tools will elevate the utility of GOS to a new level so that it is not only searching, accessing, and finding data, it will be able to provide solutions to real-world management problems. Utilizing decision-support and public-domain software, these tools cannot only add significant value to GOS, but in the long term reduce the overall development, licensing, and maintenance costs through the enterprise by re-engineering and sharing tools from one community of practice to another.

The Marine Realms Information Bank—A Coastal and Marine Digital Library

Frances L. Lightsom[1] and Alan O. Allwardt[2]

[1]U.S. Geological Survey Woods Hole Science Center, Woods Hole, Mass.
[2]U.S. Geological Survey Pacific Science Center, Santa Cruz, Calif. (aallwardt@usgs.gov)

The Marine Realms Information Bank (MRIB) is a digital library that classifies, integrates, and facilitates access to free online scientific information about oceans, coasts, and coastal watersheds, as well as the people, techniques, and organizations involved in coastal and marine science. The significance of the MRIB project lies in

(1) The utility of the digital library. MRIB provides access to Websites, full-text reports, maps, and downloadable data. The search interface accommodates three strategies: topical searching, using a faceted classification with 12 high-level categories; spatial searching, by map or gazetteer; and keyword searching.

(2) Implementation of the distributed geolibrary concept. MRIB is a gateway to information resources distributed across the Internet on many different servers, and these information resources are georeferenced by coordinates and place name.

(3) Custom entries into subsets of the MRIB database. This customization provides specialized digital libraries for particular regions or topics, and has been successfully implemented in pilot projects for Monterey Bay, California (regional focus), and coastal-change hazards (topical focus).

(4) Modular software architecture. This architecture can be used to create digital libraries for other disciplines. With appropriate modifications, the MRIB software could easily accommodate geospatial information from a wide range of natural and social sciences.

MRIB is a cooperative project of the U.S. Geological Survey (USGS) Coastal and Marine Geology Program (CMGP) and the Woods Hole Oceanographic Institution and can be found online at http://mrib.usgs.gov/. The two customized MRIB interfaces are components of the CMGP Knowledge

Bank and also are online: the Monterey Bay Science Digital Library, available at http://mrib.usgs.gov/mbs/; and the Coastal Change Hazards Digital Library, available at http://mrib.usgs.gov/cch/.

U.S. Geological Survey Science Topics Index and Supporting Infrastructure

Peter N. Schweitzer[1]

[1]U.S. Geological Survey, Reston, Va.

Science Topics, a new component of the U.S. Geological Survey (USGS) home page, is a browsable index of web resources intended for the public. It augments traditional search and site-specific browse interfaces in concert with the new "USGS by State" and "About USGS" sites and is specifically intended to work alongside those facilities.

This infrastructure is specifically intended to help people outside the USGS find information on USGS websites without specific knowledge of the organizational structure and operations. The interface design is based upon the idea that finding information cannot be separated from understanding. In the process of browsing or searching, a user is assisted by viewing relationships among scientific concepts and thereby learns while searching.

Underpinning this interface are several controlled vocabularies structured as formal thesauri and authority lists; a catalog of web resources appropriate for the intended audience; and software used to create, review, and modify both the controlled vocabularies and the catalog records. The system serves as an example of the reapplication of classic library methodologies in a web setting. LAMP is the development environment; the web browser provides the user interface.

Federal Information and Investment Programs

C. Ferrigno[1], D. Kimball[1], A. Berger[1], and J. Snoich[1]

[1]U.S. Geological Survey, Reston, Va.

The U.S. Geological Survey (USGS) participates in many Federal information and investment programs including Records Management, The Privacy Act, the Federal Rehabilitation Act Section 508, and Capital Planning and Investment Control (CPIC).

The Records Management Program is tasked with ensuring that accurate and complete documentation of the policies, procedures, functions, organization, transactions, and science of the U.S. Geological Survey (USGS) is maintained. All USGS records, regardless of form or media, are managed in accordance with regulations, from their initial creation, to maintenance and use, to their final disposition. The Records Management Program identifies USGS records through issuance of a general records schedule, which covers bureau administrative records and mission-specific (scientific) records schedules. Adequate

safeguards are established to prevent unauthorized access, removal, destruction, or loss of USGS records. The Records Management Program establishes procedures for organizing bureau files to ensure that only essential records are maintained and provide for quick and easy retrieval of records, easier identification and retention of records of archival value, and timely disposition of short-term or temporary records. The Records Management Program ensures that all USGS records appraised by the Archivist of the United States as having permanent value are forwarded in a timely manner to the National Archives and Records Administration (NARA) for preservation.

The Privacy Act of 1974 (5 U.S.C. 552a) regulates the collection, maintenance, use, dissemination, and disposal of information on individuals that are maintained in systems of record. Privacy Act requirements apply to information on individuals and information in identifiable form. The Privacy Act prohibits the collection of personal information that has not been authorized. All Federal employees who handle information on individuals, collect and file information by name, and manage a database with information on individuals are responsible for complying with the requirements of the Privacy Act.

Section 508 of the Federal Rehabilitation Act requires that any Electronic Information Technology produced, procured, maintained, or used by any Federal agency must be compliant with technical standards and procurement regulations as mandated by Section 508. For any member of the public seeking information from a Federal agency, employees must provide access to people with disabilities in a comparable manner to those who do not have disabilities. Section 508 regulates software applications and operating systems; Web-based information or applications; telecommunication products; video and multimedia products; self-contained, closed products (e.g. information kiosks, calculators, and fax machines); and desktops and portable computers. Section 508 is enforced on the USGS Web by use of an enterprise Web tool that monthly scans every registered Website for Section 508 compliance and sends reports to Web administrators who repair and republish the corrected sites.

Capital Planning is a systematic approach to managing the risks and returns of information technology (IT) investments for a given project. CPIC is a structured and integrated approach to managing IT investments and ensuring that bureau IT investments align with the mission of the U.S. Department of the Interior (USDOI) and support its business needs while minimizing risks and maximizing returns throughout the life cycle of each investment. The emphasis is placed on achieving a desired business outcome. Status of USGS IT investments are identified, measured, and reported to USDOI and Office of Management and Budget (OMB) in three categories: major IT investments (greater than $5M annual cost); non-major IT investments (less than $5M annual cost), and IT infrastructure (Communications Services, Computing Services, Electronic Work Environment, and Cross-Cutting). CPIC relies on systematic selection, control, and on-going evaluation processes to ensure that the objectives of each investment are met efficiently and effectively.

U.S. Geological Survey Enterprise Architecture—Utilization of the Data Reference Model to Improve Data Sharing

Raymond C. Obuch[1] and Stuart Doescher[1]

[1]U.S. Geological Survey, Reston, Va.

One component of the U.S. Geological Survey (USGS) Enterprise Architecture (EA) effort involves collaboration with the other bureaus in support of the U.S. Department of the Interior (USDOI) EA Program. USGS is an active participant on the USDOI Data Advisory Committee (DAC). The DAC works to promote more effective and efficient handing of USDOI data assets, including both science and business information. The ability to effectively and efficiently find, obtain, and use these data and information is a major performance metric for the DAC. A current inventory of data holdings and the computer systems and subsystems that house and deliver data and information is an essential “best practice” for data delivery and management. The DAC, in harmony with the Federal Enterprise Architecture, has developed a Data Reference Model (DRM) that provides a standard by which data can be described, categorized, and shared.

Classifying and organizing data using taxonomy such as the DRM's Subject Area and Information Classes system will aid in data discovery throughout the USDOI. The USDOI is implementing a Data Stewardship Program that will create an organizational structure to facilitate the inventory, organization, and classification of data and information in accordance with the DRM. USGS is now investigating how best to implement a Data Stewardship Program within the Bureau.

Summary: Effective data-resource management will facilitate data discovery and data sharing between and among communities of interest.

A Global Organism Detection and Monitoring System

Catherine S. Jarnevich[1], Thomas J. Stohlgren[1], James J. Graham[2], and Gregory J. Newman[2]

[1]U.S. Geological Survey Fort Collins Science Center, Fort Collins, Colo.
[2]Colorado State University, Fort Collins, Colo.

Efficient ecological monitoring of the many costly and harmful invasive organisms now widespread throughout the United States requires real-time information readily available to land mangers and the public. However, no central repository designed to provide such access currently (2006) exists. Individual organizations typically collect and manage their own information using their own data-management systems; there is no integration of these disparate data sets. To meet this need, the U.S. Geological Survey (USGS) National Institute of Invasive Species Science created a Global Organism Detection and Monitoring system (GODM) to track invasive species as their ranges change either by spreading across the country or contracting from control efforts. This system involves an integrated suite of tools that enables users to seamlessly take data collected in the field to a centralized database, to an integrated analysis and modeling service built in collaboration with the National Aeronautics and Space Administration (NASA), directly to a web-service enabled website accessible to anyone with a web connection. The GODM system allows users to upload field data using text files, shapefiles, or pre-formatted formats such as the GODM system's own field tools (e.g. custom PALM applications for weed mapping and vegetation surveys) and, in the future, Weed Information Management System (WIMS) exports. Once in the database, the integrated disparate data sets can be used to create distribution maps, watch lists, species lists for areas, and, in the future, predictive models using NASA's Invasive Species Forecasting System, among other things. These tools should help resource managers in their effort to prevent the spread of non-native, invasive species.

Development and Utilization of a National Geochemical Database

Steven M. Smith[1] and David B. Smith[2]

[1]U.S. Geological Survey, Mineral Resources Program, Denver, Colo. (smsmith@usgs.gov)
[2]U.S. Geological Survey, Mineral Resources Program, Denver, Colo. (dsmith@usgs.gov)

The Geologic Discipline of the U.S. Geological Survey (USGS) has had a long history of collecting and analyzing earth materials. Samples of rock, mineral, soil, sediment, water, vegetation, and even animal tissue have been analyzed while researching topics in mineral deposits, petrography, mineralogy, geologic mapping, alteration, geochronology, mineral resources, energy resources, health, and the environment. In the 1960's, with the advent of computers, sample information and associated analytical data began to be saved in database files for possible future use. As time passed and political administrations, philosophies, laboratories, analytical methods, and personnel changed, this early database structure evolved into a series

of incompatible geochemical databases within the USGS. Potential users needed to become familiar with the intricacies of USGS thought, projects, and methods just to find, retrieve, and understand the data.

The National Geochemical Database (NGDB) project began with the ambitious goal of taking the historical USGS geochemical databases plus the inherited geochemical database from the U.S. Department of Energy's National Uranium Resource Evaluation (NURE) program and combining them all into a single format. This single database also was envisioned to be compatible with new analytical data being produced in USGS laboratories. Although simple in concept, this process became increasingly complex as the different databases were merged and as political administrations, philosophies, laboratories, analytical methods, and personnel continued to change.

Although not yet complete and fully functional, the NGDB currently (2006) stores almost 2 million sample records and 39 million analytical determinations. A conservative estimate of the value of these data, based upon costs to reacquire the information, ranges from 1 to 1.5 billion U.S. dollars. In addition to the database, the USGS has archival splits of most of the samples making up the NGDB, thus affording the opportunity for reanalysis to obtain data for additional elements or to determine concentrations by improved analytical techniques. Efforts are underway to increase, verify, and correct data in the NGDB and to make these data readily available on CD-ROM, DVD, and through on-line access. Selected subsets of the NGDB are currently being served on-line and with no charge at http://tin.er.usgs.gov. The National Geochemical Database project offers valuable lessons on how to and how not to handle geochemical databases.

The South Florida Information Access (SOFIA) System

Heather S. Henkel[1]

[1]U.S. Geological Survey Florida Integrated Science Center, St. Petersburg, Fla.

The South Florida Information Access (SOFIA) system was created by the U.S. Geological Survey (USGS) in 1995. Its mission is to provide easy access to information about research projects and products generated as part of the USGS South Florida Priority Ecosystem Studies (PES) Program and other Federal, State, and local science providers. SOFIA provides this service by integrating information systems and tools enabling efficient storage, organization, and search and retrieval of scientific information about the south Florida ecosystem. SOFIA was designed to benefit three major user groups: USGS program managers and scientists working with the South Florida PES Program, managers and scientists working for other organizations involved with Everglades restoration, and members of the public interested in USGS research and (or) the science behind the Everglades restoration effort.

SOFIA is an evolving and dynamic system that builds on the ever-increasing sophistication of new information technology. The current architecture consists of three integrated components: website, data, and metadata. The SOFIA website (http://sofia.usgs.gov/) contains links to project descriptions, proposals, publications, data (through links to our data exchange site), metadata, presentations, and contact information, as well as general interest items, such as photographs and posters. The SOFIA site also is a portal through which you can access our extensive data sets and Internet map server (IMS).

Data is served by two mechanisms on the SOFIA website. The Data Exchange (http://sofia.usgs.gov/exchange/) provides access to files organized by project. The projects are further organized using six primary themes: biology, chemistry, ecology, geology, hydrology, and mapping. The second mechanism of serving data is through a web-based map server. The map server, which is being developed using ArcIMS software, will provide a means of accessing information stored in an Oracle database and the SOFIA data exchange website through a geospatial query.

Large amounts of data have been collected by USGS personnel in south Florida. With good, FGDC-compliant metadata, the data are available to a much wider set of customers through web-based queries. The SOFIA website has all the available metadata accessible by several methods. There is a navigation button for Metadata and each project home page has a listing for its associated metadata for the project and for the data. Work is continuing on updating the metadata for completed projects and for remaining data sets that do not yet have metadata.

myUSGS Portal Pilot Project

Sky Bristol[1]

[1]U.S. Geological Survey, Denver, Colo.

The myUSGS pilot project has been created to explore and test new methods of enabling teams of scientists to more effectively communicate and to share and manage information and data generated in the course of planning and conducting a specific study or scientific initiative. The project is being conducted by the Geospatial Information Office (GIO), in collaboration with a number of key science program partners, including

- Mancos Shale Landscapes Integrated Science Project
- Integrated Landscape Monitoring Science Thrust (Great Basin Project, Puget Sound Project, Prairie Potholes Project, Lower Mississippi Valley Project)
- Fire Science Thrust (Colorado Front Range Project, Western Montana Project, Great Basin Project)
- Water Availability Science Thrust
- Landslides and Debris Flow Science Thrust

The objective of the pilot project is to work with each of these science communities to create a web-based "science project portal" consisting of an interface to a set of capabilities and tools tailored to the needs of each project. These can include tools that can help the team members collaborate and share information among themselves, as well as tools that support inventory, integration, and use of existing data sets. The portals will use common and consistent technology so that each Science Thrust team doesn't have to build a site from scratch and will integrate features that are generally interoperable across different portals. The portals will be customizable, flexible, and modular in design so that as new tools are made available, they can be easily added to any portal.

Geographic Information System for the Gulf—ADS40 Imagery on Lidar for Hurricane Katrina

David Greenlee[1]

[1]U.S. Geological Survey, Sioux Falls, SD

A Geographic Information System "(GIS) for the Gulf" (GFG) demonstration is available that takes advantage of the Geospatial One-Stop portal and geographic information that was collected for hurricane affected areas in the Gulf coast. The GFG database was assembled collaboratively by the U.S. Geological Survey (USGS) and Environmental Systems Research Institute (ESRI), using a data-model concept called GIS for the Nation. In the GFG instance, reference data (i.e., framework, foundation, ...) was gathered,

along with pre- and post-Katrina imagery and Federal Emergency Management Agency (FEMA)-contracted photo interpretations of the damage classes. The GFG database covers a range of scales from regional to local to neighborhood and includes data from *The National Map*, parcels collected from counties and parishes, and satellite and aircraft imagery.

StreamStats: A Web Application for Streamflow Statistics and Basin Characteristics

Kernell Ries[1]

[1]U.S. Geological Survey, Baltimore, Md.

StreamStats is a Web application (http://water.usgs.gov/osw/streamstats/) that allows users to obtain streamflow statistics, drainage-basin characteristics, and other information for user-selected sites on streams. StreamStats users can choose locations of interest from an interactive map and obtain information for these locations. If a user selects the location of a USGS data-collection station, StreamStats will provide previously published information for the station from a database. If a user selects a location where no data are available (an ungaged site), a Geographic Information System (GIS) program will estimate information for the site. The GIS program determines the boundary of the drainage basin above the site, measures several physical characteristics of the drainage basin, and solves regression equations to estimate streamflow statistics for the site. In the past, it could take an experienced person more than a day to estimate streamflow statistics for an ungaged site. StreamStats reduces the effort to only a few minutes.

Collaboration in the Development of the Bird Banding Laboratory Database System

Kevin Laurent[1]

[1]U.S. Geological Survey, Reston, Va.

Collaboration among participants in an information-technology development project has always been important. A poor collaboration and communication strategy has been cited many times as the major reason for the failure of a development project. What do you do when you have diverse stakeholders; national and international attention; a project already way-behind schedule; and a dedicated, energetic staff? Collaborate! Collaboration was employed in the recent development of a new Grand Design database project for the Bird Banding Laboratory, which turned a project headed for failure into a successful deployment.

Acknowledgments

Thanks to Tom Gunther, Denise Wiltshire, David Govoni, Bill Miller, and Fran Lightsom for their guidance, comments, and suggestions. Special thanks to Alan Allwardt and Chris Polloni for their reviews.

Appendix I.

Agenda
U.S. Geological Survey
Scientific Information Management Workshop

March 21–23, 2006
USGS National Center
Reston, Virginia

Pre-Session (Monday, March 20, 2006)

6:30 pm – 8:30 pm Pre-session for workshop organizers and presenters at the Marriott Dulles Suites at Worldgate — Salon 1

Day 1 (Tuesday, March 21, 2006)

7:30 am – 8:30 am Registration

8:30 am – 9:00 am Welcome to the workshop
Overview of workshop objectives and outcomes

9:00 am – 9:30 am Director's Charge for Workshop

9:30 am – 10:30 am Presentations I — Scientific Information Management: Practices, Challenges, and Opportunities

Forum for presentations by natural and information scientists and information managers on the diversity of approaches to information and data management by the programs; existing projects and data-management systems; and challenges to meet mandates, mission needs, and cooperator requirements.

- Data-Management Challenges for the USGS Volcano Hazards Program — Dr. Peter Cervelli and Dr. Jim Quick (Presenter) (USGS Volcano Hazards Program)
- Things the LTER Learned Managing Long-Term Data Sets — Dr. Inigo San Gil (Long Term Ecological Research Network and NBII program office)

10:30 am – 10:45 am Break

10:45 am – Noon Presentations II — Scientific Information Management: Practices, Challenges, and Opportunities (continued)

- The IPANE Program: The Synergism of Science and Public Involvement — Dr. Leslie J. Mehrhoff (Invasive

Plant Atlas of New England, University of Connecticut)

- Making Sense of it All: An Ecologist's Perspective on National Databases and Data Analysis Tools in the NAWQA Program — Dr. Thomas F. Cuffney (Research Ecologist, National Water-Quality Assessment Program, USGS North Carolina Water Science Center)

Noon – 1:00 pm Lunch

1:00 pm – 2:00 pm Keynote Address — Dr. Etienne Wenger, Learning for a Small Planet

Dr. Wenger is a leading expert on communities of practice. He is the founder of Learning for a Small Planet, an investigation into fostering learning institutions, and former Research Scientist at the Institute for Research on Learning, where he developed his learning theory centered on the concept of community of practice. For the last 6 years, he has been helping organizations develop and implement knowledge strategies based on communities of practice.

2:00 pm – 3:15 pm Panel I — Communities *in* Practice

The community of practice approach as applied and practiced in other organizations

- Etienne Wenger (Convener)
- Bill Knapp (USFWS)
- Susan Mockenhaupt (U.S. Forest Service)
- Laure Wallace (USGS Office of Human Resources)
- Mark Youman (ICF Consulting)

3:15 pm – 3:30 pm Break

3:30 pm – 4:30 pm Open Discussion — The Intersection of Communities of Practice and Scientific Information Management at the USGS

A dialogue between presenters, panelists, and the audience

4:30 pm – 4:45 pm Closing remarks

7:00 pm Evening Reception at the Marriott Dulles Suites at Worldgate

Day 2 (Wednesday, March 22, 2006)

8:30 am – 9:00 am Overview of Day 2 activities

9:00 am – 9:30 am Words Matter: Developing a Common Vocabulary for Common Understanding

Introduction to a group activity for reviewing definitions of concepts,

standards, and processes to arrive at common terminology for information and data-management practices.

9:30 am – Noon Open Session — Demonstrations and Posters

Focus on information-management issues as well as tools and systems used for information and data management that may be shared among programs.

Noon – 1:00 pm Lunch

1:00 pm – 1:15 pm Overview of community strategy and charge to communities

1:15 pm – 3:15 pm Community Building Sessions I

- Archiving of Scientific Data and Information (John Faundeen)
- Digital Libraries (Steve Shivers)
- Field Data for Small Research Projects (Fran Lightsom)
- Knowledge Capture (Laure Wallace)
- Portals and Frameworks (Sky Bristol)
- Scientific Data from Monitoring Programs (Paul Geissler)

3:15 pm – 3:45 pm Break

3:45 pm – 4:30 pm Report-outs from Community Building Session I

4:30 pm – 5:30 pm Birds–of–a–Feather sessions

Day 3 (Thursday, March 23, 2006)

8:30 am – 8:45 am Recap of Day 2 community building sessions
Overview of Day 3 activities

8:45 am – 10:45 am Community Building Sessions II

- Database Networks (Chris Polloni)
- Emerging Workforce (Pamela Malam)
- Knowledge Organization Systems and Controlled Vocabularies (Peter Schweitzer)
- Large Time Series Data Sets (Harry House)
- Metadata (Sharon Shin)
- Preservation of Physical Collections (Martha Garcia)

10:45 am – 11:00 am Break

11:00 am – Noon Report-outs from Community Building Sessions II

Noon – 1:00 pm	Lunch
1:00 pm – 2:30 pm	Panel II — Perspectives on the Workshop: Impressions, Issues, and Implications • Martha Garcia (USGS-BRD Priority Ecosystems Initiative) (Convener) • Mark DeMulder (USGS-GIO Science Information and Education Office) • John Faundeen (USGS-GIO Archivist) • Linda Gunderson (USGS-GD Acting Associate Director for Geology) • Fran Lightsom (USGS-GD Coastal and Marine Geology Program) • Peter Lyttle (USGS-GD Cooperative Geologic Mapping Program) • Karen Siderelis (USGS-GIO Associate Director for Geospatial Information)
2:30 pm – 2:45 pm	Break
2:45 pm – 4:30 pm	Open Discussion — Future Directions and Next Steps
4:30 pm	Adjourn

Appendix II. Information-Management Principles for the USGS Scientific Information Management Workshop

William G. Miller

Introduction

An important aspect of joint activities is a common understanding of at least three things: (1) a shared vision of what is to be done, (2) the meaning of terms used to describe the activity, and (3) the principles to be followed to accomplish the activity. The principles listed below were intended to stimulate discussion regarding the principles and practices that should guide information management at the USGS. Some of the entries are more obvious than others. Each principle has a stated rationale and some of the implications of its implementation to provide a basis for discussion. Principles, being over-arching guiding ideas, are broad and few in number. In fact, only eight are proposed.

Each attendee was encouraged to make comments by placing sticky notes on posters associated with each principle. The results were collected and are presented later in this volume. The intent of this exercise was to start a discussion of these principles within the community of information managers present at the workshop.

Principle 1: Data and information are different.

Rationale

1. Data are the carriers of information. Data may be transformed into many formats but the information carried by them remains the same during ideal transformations.

Implications

1. The economic value of information and data can be computed differently.
2. The mechanics of managing data is different than managing information.
    a. Data are stored, retrieved, and transmitted with no attached context.
    b. Information requires a context (metadata) to be useful.
    c. Data can be created, transformed, and destroyed.
    d. Information is not destroyed by use.
    e. Corrupt data cannot effectively carry information.

Principle 2: The quality of information is of prime importance.

Rationale

1. Effective decisions are made with good quality information. Bad information causes bad results.
2. Information-quality standards are required by statute and policy.

Implications

1. What is meant by information quality must be better understood, defined, and documented.
2. Quality standards used must be documented. (Required by policy and law)
3. There must be an effective method for managing the quality of both structured and unstructured data.
4. Owners of information, or possessors of knowledge, must document the purpose (including its characteristics) that it was collected for before it is shared with others.
5. There must be agreed standards, processes, and appropriate tools to ensure that the appropriate version of information and knowledge is being used.
6. Principle 5 becomes important to the implementation of this principle.

Principle 3: Information will be made available to all at the lowest cost.

Rationale

1. Exchange of scientific information is required by the scientific method.
2. Unrestricted access to information is a public good.

Implications

1. Data may have a cost but information does not. Fees are charged only for conveyance.
2. Some data are restricted from disclosure. Mechanisms must be available to identify and protect them. The fine point here is the protection from conveyance of information.

Principle 4: Data will have clearly defined ownership.

Rationale

Note: This is one of the differences between data and information. Information cannot be owned.

1. Credit for the collection and preservation needs to be given.
2. Data management requires planning and stewardship to achieve the goals of Principle 2.

Implications

1. Data ownership conveys responsibility.
2. Planning for the preservation of data will be required to discharge the associated responsibilities.
3. Ownership of data may change over time to accomplish Principle 6.
4. Data owners will incur costs.

Principle 5: Use of information will be based on common understanding of its semantics.

Rationale

1. The meaning of information is context dependent.
2. For information to be interoperable among systems, the meaning associated with it must be agreed upon.
3. Common semantics is a cornerstone of communication and sharing.

Implications

1. Investments must be made in creating machine-readable consensus documents (ontologies) that define terms and the contexts (namespaces) for which they are valid.
2. The ontologies must be readily available in a standard form.
3. Software must utilize the ontologies to inform the user of the data sets of its interoperability status.

Principle 6: Information will be maintained for future uses.

Rationale

1. To eliminate the redundant cost of capturing the same data repeatedly
   a. A monitoring activity will make repeated measurements or observations to gain statistical clarity or for change detection.
   b. Only truly redundant data creation should be eliminated.
2. To ensure that decisions are made using the most accurate information possible

Implications

1. Agreements on standards and methods are required.
2. Resources need to be made available to accomplish this.
3. The information must be retrievable at a later time (potentially decades).
4. The data carrying the information must be readily transformable.

Principle 7: Tools and techniques for data and information management will be shared as widely as possible.

Rationale

1. Direct costs must be minimized.
2. Training costs must be minimized.
3. Responsibilities need to be met.
4. Widest availability of information is desired.

Implications

1. Interoperability of the tools and techniques are more important than simple operability because tools, techniques, and platforms change over time.
2. Principle 5 becomes important.

Principle 8: All data and information will be managed in compliance with legal obligations.

Rationale

1. This is a basic tenet of civilized society. We, as Government employees, have a duty to comply with legal requirements.

Implications

1. Staff must be competent to use and manage data, information, and knowledge in compliance with legal statutes and other endorsed codes of practice and instruments.
2. USGS must organize its data flexibly so that it can meet legal requirements quickly and at minimal cost.
3. There will be costs over and above basic project costs to do this. Project managers must take these added costs into account.
4. The status of data may change and expenditures made to manage it according to the then current standards

When data owners become aware of changes in statutory, regulatory, or policy requirements they should alert management of the implications including long-term costs and programmatic effects.

The Results of the Principles Discussion

During the workshop, a set of posters were placed on the wall for people to comment on the proposed principles. The results of those comments are as follows:

Principle 1: Data and information are different.

- [Principle 1 provides] one [of many] perspective[s] on data versus information.
- Knowledge is different from both data and information.
- Think of mining [as] an analogy: “data” equals the ground; “information” equals coal extracted from the ground; “knowledge” equals energy extracted from consumed coal.
- The difference doesn't matter.

Principle 2: The quality of information is of prime importance.

- Who determines the meaning of quality? The producer? The sponsor? The consumer?

- Required quality (e.g., standards) versus desired quality: who or what decides?
- A quality has a cost that must be balanced against other requirements.

Principle 3: Information will be made available to all at the lowest cost.

- These principles are USGS oriented. A private sector information provider would differ.

Principle 4: Data will have clearly defined ownership.

- To use a software analogy: Data should be open access, not proprietary. Therefore, data should be owned by all. Maintenance, stewardship can be [an] individual responsibility.
- Most USGS data are paid for by taxpayers. Are [these data] reported from the owner?
- Contrast owning an economic right to control data versus stewardship

Principle 5: Use of information will be based on common understanding of its semantics.

- No, retrieval will be. Use is a different issue.

Principle 6: Information will be maintained for future uses.

- Information has a lifecycle and value [that] often declines below certain thresholds.
- Data should be maintained for future uses.

Principle 7: Tools and techniques for data and information management will be shared as widely as possible.

- We need a tools/techniques library.
- Should the USGS have a warehouse of old technology to read older data, i.e. punch cards?
- Don't reinvent the wheel! Share! Use successful data management/distribution techniques USGS-wide. For example, Denver uses are ArcIMS quite successfully for products with a GIS component. What does Reston use?

Principle 8: All data and information will be managed in compliance with legal obligations.

- Not much to discuss here, unless “policy" is open to interpretation.
- Bureau wide (across *all* disciplines)
- Owners must remember their legal requirements as stated in the federal records act—manage your records through creation, use, maintenance, and disposition.

It seems clear from these comments that there are many perspectives on the issues of information management. These differences should be explored further for the community to work more efficiently and effectively.

Appendix III. Community Building Session Descriptions

Archiving of Scientific Data and Information

Archiving is generally considered the last step in the records-management lifecycle of creating, using, and maintaining records of value. Common USGS perceptions of archiving have a common theme, i.e. the implied preservation of our scientific work for future generations of scientists. Adequately preserving our work has become a large challenge owing to how programs and projects end and new ones start. The pace of obsolescence of hardware, software, firmware, operating systems, and media compound the problem. Finally, the lack of specific, assigned responsibility for preserving our work long-term coupled with inadequate resources to carry out the task leaves us wanting in terms of being able to provide adequate preservation and access to our scientific data for researchers, policymakers, and the public. This CoP will seek to

1. Identify individuals interested in seeking solutions to our archiving challenges
2. Attempt to list specific USGS scientific data requiring rescue efforts
3. Develop strategies to begin addressing the identified needs
4. Discuss archive-media options
5. Explore scientific-records appraisal approaches
6. Brainstorm off-site records-storage options

The most important objective will be to identify those individuals willing to discuss these and other topics raised during the workshop through ongoing forums. Sharing our experiences, tools, and services will aid the Bureau in beginning to address the problem of preserving our science.

Digital Libraries

Scientific research requires high quality library support in order to succeed. Library science is undergoing a digital revolution that will have an effect on everyone in the USGS, yet the concept of the "digital library" as it relates to USGS is not clearly defined. To some it may refer only to the digitized holdings from a traditional library collection. In reality, it is most likely much more than that. Any collection of digital information along with the tools and support staff that make that information useful to its users could be considered a digital library. The types of information included in a digital library are only limited by the imagination. The overriding factor in moving from traditional libraries to the digital world is cost. This community can search for common ground in the efforts to develop digital libraries across the USGS. Developing a consensus on what constitutes a digital library within the USGS would benefit all parties involved, as would the bringing together of library professionals and those who have digital library needs in their programs. Gathering common requirements and developing some uniform strategies to meet those needs might be an achievable goal. This community may develop into a united pool of knowledge and current resources within the Bureau that could lead to appreciable cost savings as developments continue.

Field Data for Small Research Projects

USGS research scientists carry out small-scale field projects to study particular processes or local areas, projects which are not designed to contribute to a large-scale or long-term database. Because of their limited scope, duration, staffing, and funding, these projects can benefit from assistance with USGS fundamental science practices, particularly in archiving data beyond the duration of the project, and identification of useful Bureau-wide standards and data-access systems. Possible approaches to this set of issues are provision of expert consultants and advisors, engagement of project scientists and technicians in developing standards and handbooks, training of project personnel, and easy-to-use tools and repositories. This community might break into sub-groups that address particular issues or particular approaches.

Knowledge Capture

Knowledge capture is a critical step in turning data and information into reusable knowledge. While definitions differ, most agree that knowledge capture is an essential step in knowledge management—the process by which we create, identify, and distribute organizational knowledge to those who need it. It is not about capturing knowledge for knowledge sake. The goal is to increase the ability of people to share and codify best practices and information and create new knowledge. Power lies in the ability to share knowledge in a way that leads to continual learning in the organization and the synergistic creation of best practices for larger organizational benefit.

While technology can provide the means to organize and quickly access knowledge, critical knowledge also can be captured and shared in a variety of other ways. Mentoring, CoP, and joint problem solving among experts and novices are excellent techniques for sharing knowledge and best practices on how to approach complex problems.

The USGS faces the loss of critical knowledge through retirements and attrition unless we develop methodologies to preserve and enhance this knowledge. The dispersed nature of work makes a case for processes to share best practices and develop new understandings through groups such as CoP. The amount of time and resources available to experience and acquire personal knowledge is limited and makes a case for captured and shared understanding. Employees at all levels are in contact with the same customers necessitating the need to have information about these interactions shared rapidly to leverage that knowledge to serve those customers better.

In rare moments, there are instances where people with great ideas and knowledge come together and are able to make something even better. Perry Glasser, CIO Magazine said, "… the great things often begin at those fortunate moments when people with knowledge and vision pool their dreams." We can no longer depend upon the serendipity of those fortunate moments. Knowledge capture and management must be systematically pursued in the culture, shared, and used to help us adapt and grow to ensure our continued scientific excellence.

Portals and Frameworks

Increases in the speed at which scientific information can be produced and made available have resulted in vast quantities of data and tools that are potentially useful to any project. The process of determining utility and putting those data and tools into use can be extremely cumbersome and time consuming. Some combination of information standards, proven methodologies, and interoperable tools must be brought together into one or more frameworks that can be applied to various scientific-research questions. One such framework is found in portal technologies that provide a Web-based system of interrelating applications and information sources. The continuing evolution of the Web, with the proliferation of consumable Web services (geospatial and otherwise) and robust client applications that can tap into these resources, begin to enable a world where scientific data can be published in a readily accessible form and quickly plugged into countless applications within a standards-based framework of tools and capabilities. This community will examine the current uses and future potential of information frameworks in the USGS. While portals and other technologies may form viable tools today, the information produced through these applications must exist beyond any one software iteration. The community will take the big-picture view to address standards within these frameworks that will stand the test of time.

Scientific Data from Monitoring Programs

The USGS has a special mission to collect, maintain, and distribute long-term monitoring data for critical geological, hydrological, and biological processes. Monitoring programs face information-management challenges that result from the need for consistent standards and protocols for a large variety of measurement situations; consistency through organizational, personnel, and technological changes; and continuous provision of a significant body of real-time information to a diverse set of customers. These issues require a combination of technical, information-science, and managerial approaches. This community might exchange information about advances in scientific- and information-technology techniques for continuous and consistent scientific sampling and (or) reporting; advances in web services and interaction design; management strategies for training and maintaining a distributed network of skilled technicians and volunteers; and (or) organizational strategies for coordinating among cooperating agencies and organizations.

Database Networks

The USGS collects, stores, and serves scientific data and information in a large number of databases of varying size, temporal and geographic scale, and complexity. They are widely used by researchers, managers, and others for visualization, hypothesis testing,

modeling, and other forms of analysis. There is a growing desire to access and use these databases in combination, particularly in real-time, on-demand, and dynamic ways. While various "middleware" tools exist that can help support this type of access, there are many questions regarding a broadly useful, reliable, and sustainable approach for the USGS. Solutions may include commercial or open-source software, better exploitation of metadata, "cook-off" comparisons of alternative approaches, and broader adoption of standard tools (e.g., XML or web services).

Emerging Workforce

The effective and efficient delivery of USGS-mission programs depends heavily upon an organization's ability to recruit, retain, develop, and deploy a workforce. Human capital (including employees, contractors, students, and other affiliates) represents the single largest investment the USGS makes each year. Effectively utilizing this investment requires an understanding of and anticipating both workforce and workplace trends.

Workforce and workplace trends are characterized under the broad topical areas of demographics, economics, employment, globalization, politics, science and technology, and society. Examples include issues relating to an aging workforce; rising health-care costs; and advanced uses of technology, safety and security, and corporate ethics.

Raising awareness of key workforce and workplace trends to the attendees of the SIM Workshop will facilitate discussion on the development of a CoP approach to meeting the challenge of making SIM easier.

Knowledge Organization Systems and Controlled Vocabularies

Science can be communicated only by using a common language to represent its concepts. The language must be used with consistent application and meaning to be effective and efficient. Controlled vocabularies, along with more complex structures such as topic maps and ontologies, provide ways to explain the relations among scientific concepts we use to carry out scientific research. What USGS needs from this field are mechanisms for representing, describing, using, and presenting KOSs among ourselves and for the public. The KOSs themselves are used to categorize information resources from broad-scale (for example, topical descriptions of research data collections) to fine-scale (for example, identifying the function of metadata elements or simply snippets of text).

Large Time Series Data Sets

Many USGS science programs collect data on a regular basis, creating temporal or time series data collections. They vary in the frequency with which data are collected (from yearly to nearly continuous), the range of data collected (from a single measure to a suite of environmental variables), and the scale of data collected (from national to local). Various tools exist (and some are implemented) to assist in the unique requirements related to time series data collection, means to transmit to and read from centralized databases, the serving of data in near real-time, and time series analysis. Sharing the knowledge embodied in these efforts, including factual information, "real world" experience, training and education options, future development plans, and the evolving application needs of science programs are potential topics for this discussion and a possible future CoP.

Metadata

The USGS, as well as other government agencies, is awash with vast information resources. Effective resource discovery and access can be elusive and frustrating for USGS personnel as well as the public. Inventorying and documenting resources through the use of metadata assists information management and may lead to improved resource discovery and access. Metadata has been implemented for resources such as geospatial and scientific data sets and collections, web content, museum and voucher collections, methodologies and protocols, digital objects (e.g., photographic images and video and audio streams), modular training-management systems, and document archives. This CoP could become a focal point to exchange best practices, share evolving technologies and techniques, collaborate on and coordinate information-management activities, and harness the vast information resources within the USGS disciplines.

Preservation of Physical Collections

The Administration's Fiscal Year 2007 Research and Development Budget priorities memorandum identified Federal scientific collections as an area requiring special Agency attention. As a result of the memo, an Interagency Working Group on Scientific Collections (IWGSC) was established in December 2005 to access the priorities and stewardship of scientific collections and to develop a coordinated strategic plan to identify, maintain, and use Federal collections. Initial information regarding scientific collections will be gathered through a web-based survey that will be used to outline strategies for better coordinating collection activities and increasing awareness of the collections. This community could play a pivotal role in shaping effective policy and technical approaches to collections preservation by such means as provision of expert advice; engagement of collection managers, scientists, and technicians in developing shared standards and best practices; and encouraging and supporting development of easy-to-use discovery, management, and retrieval systems and tools.

Appendix IV. List of Attendees

Last Name	First Name	E-mail	Affiliation
Ahrendts	Susan	seahrens@usgs.gov	USGS
Allord	Gregory (Greg)	gjallord@usgs.gov	USGS
Allwardt	Alan	aallwardt@usgs.gov	USGS
Altheide	Phyllis	paltheide@usgs.gov	USGS
Balthrop	Barbara	balthrop@usgs.gov	USGS
Banowetz	Michele	michele_banowetz@usgs.gov	USGS
Best	Ronnie	ronnie_best@usgs.gov	USGS
Bier	Robert	rbier@usgs.gov	USGS
Billone	Marilyn	mabillon@usgs.gov	USGS
Bostwick	Candice	cmbostwi@usgs.gov	USGS
Brazhnik	Olga	brazhnik@nih.gov	NIH
Bristol	Sky	sbristol@usgs.gov	USGS
Broussard	Linda	linda_broussard@usgs.gov	USGS
Brown	Kim	kimbrown@usgs.gov	USGS
Bruno	Rebecca	rbruno@usgs.gov	USGS
Burress	Theresa	tburress@usgs.gov	USGS
Bushly	Thomas	tjbushly@usgs.gov	USGS
Bybell	Laurel	lbybell@usgs.gov	USGS
Callaghan	Robert	rcallaghan@usgs.gov	USGS
Campbell	Patricia	pacampbe@usgs.gov	USGS
Carswell	Bill	carswell@usgs.gov	USGS
Cavanaugh	Dan	dkcavanaugh@usgs.gov	USGS
Clines	Tom	tclines@usgs.gov	USGS
Compher	Arlene	acompher@usgs.gov	USGS
Cook	Sally	scook@usgs.gov	USGS
Coyle	David	dlcoyle@usgs.gov	USGS
Cuffney	Tom	tcuffney@usgs.gov	USGS
Cushing	Janet	jcushing@usgs.gov	USGS
Dadisman	Shawn	sdadisman@usgs.gov	USGS
Danchuk	Wendy	wdanchuk@usgs.gov	USGS
Degnan	Carolyn	cdegnan@usgs.gov	USGS
D' Erchia	Terry	terry_derchia@usgs.gov	USGS
Dietterle	Jeff	jdietterle@usgs.gov	USGS
DiNardo	Tom	tpdinardo@usgs.gov	USGS
Durant	Joye	jldurant@usgs.gov	USGS
Faries	Nancy	nfaries@usgs.gov	USGS
Faundeen	John	faundeen@usgs.gov	USGS

Last Name	First Name	E-mail	Affiliation
Ferderer	David	dferdere@usgs.gov	USGS
Ferrigno	Carmelo	cferrigno@usgs.gov	USGS
Foley	Kevin	kfoley@usgs.gov	USGS
Foulke	Donna	donna_foulke@usgs.gov	USGS
Frame	Mike	mike_frame@usgs.gov	USGS
Frank	Anthony	amfrank@usgs.gov	USGS
Freeney	Jean	jfreeney@usgs.gov	USGS
Frondorf	Anne	anne_frondorf@usgs.gov	USGS
Fuller	Kit	kitfuller@usgs.gov	USGS
Garcia	Martha	mgarcia@usgs.gov	USGS
Garrett	Lynda	LGarrett@usgs.gov	USGS
Geiger	Linda	lgeiger@usgs.gov	USGS
Geissler	Paul	Paul_Geissler@usgs.gov	USGS
Govoni	David	dgovoni@usgs.gov	USGS
Greenlee	Dave	greenlee@usgs.gov	USGS
Gunther	Gregory	ggunther@usgs.gov	USGS
Gunther	Thomas	thomas_gunther@usgs.gov	USGS
Hadley	Alexandra	ahadley@usgs.gov	USGS
Hastings	Jordan	jordan@geog.ucsb.edu	USGS
Hebenton	Tod	thebenton@usgs.gov	USGS and NBII
Henkel	Heather	hhenkel@usgs.gov	USGS
Horwitz	Lief	lief_horwitz@usgs.gov	USGS
Hothem	Larry	Lhothem@usgs.gov	USGS
House	Harry	hrhouse@usgs.gov	USGS
Hutchison	Vivian	vhutchison@usgs.gov	USGS
Jacobs	Ruth	ruth_jacobs@usgs.gov	USGS
Jarnevich	Catherine	catherine_jarnevich@usgs.gov	USGS
Kase	Kate	kate_kase@usgs.gov	USGS
Kavalek	Irena	ikavalek@usgs.gov	USGS
King	Stephen	stephenking@usgs.gov	USGS
Kirk	Keith	kkirk@usgs.gov	USGS
Klima	Karen	kklima@usgs.gov	USGS
Knapp	William	William_Knapp@fws.gov	USFWS
Kochman	Howard	hkochman@usgs.gov	USGS
Kolva	James	jrkolva@usgs.gov	USGS
Laurent	Kevin	klaurent@usgs.gov	USGS
Leake	Linda	lleake@usgs.gov	USGS
Lebing	Gerry	glebing@usgs.gov	USGS
Levine	Marc	mlevine@usgs.gov	USGS
Liberatore	Ann	aliberat@usgs.gov	USGS

Last Name	First Name	E-mail	Affiliation
Lienkaemper	George	george_lienkaemper@usgs.gov	USGS
Lightsom	Frances	flightsom@usgs.gov	USGS
Lindblom	Kathy	klindblo@usgs.gov	USGS
Liszewski	Michael	mjlisz@usgs.gov	USGS
Lofton	Ron	rlofton@usgs.gov	USGS
Lucke	Liz	liz_lucke@usgs.gov	USGS
Lyttle	Peter	plyttle@usgs.gov	USGS
Mann	Dennis	dmmann@usgs.gov	USGS
Mannstedt	Kathy	kmannstedt@usgs.gov	USGS
Marcus	Susan	smarcus@usgs.gov	USGS
Matthias	Robert	rmatthias@usgs.gov	USGS
McDermott	Mike	mmcdermo@usgs.gov	USGS
McEwen	Scott	wsmcewen@usgs.gov	USGS
Merrick	Timothy	trmerrick@usgs.gov	USGS
Miller	William	bmiller@usgs.gov	USGS
Mockenhaupt	Susan	Susan.Mockenhaupt@usda.gov	USDA Forest Service
Morris	Gene	gmorris@usgs.gov	USGS
Nandiwada	Sarojini	snandiwada@usgs.gov	USGS
Niemann	Brand	Brand_Niemann@epamail.epa.gov	USEPA
O'Connell	Jillian	joconnell@usgs.gov	USGS
Olson	Annette	alolson@usgs.gov	USGS
Ornelas	Jerry	jxornelas@usgs.gov	USGS
Payne	Rodney	rwpayne@usgs.gov	USGS
Peterjohn	Bruce	bpeterjohn@usgs.gov	USGS
Phillips	Dan	dphillips@usgs.gov	USGS
Polloni	Chris	cpolloni@usgs.gov	USGS
Pruett	Tina	tpruett@usgs.gov	USGS
Redman	Phil	pjredman@usgs.gov	USGS
Reid	Carolyn	clreid@usgs.gov	USGS
Ries	Kernell	kries@usgs.gov	USGS
Rusanowski	Chris	crusanow@usgs.gov	USGS
Russell-Robinson	Susan	srussell@usgs.gov	USGS
Sanders	Rex	rsanders@usgs.gov	USGS
San Gil	Inigo	isangil@lternet.edu	LTER Network Office
Sayer	James	jsayer@usgs.gov	USGS
Schmid	Lorna	lorna@usgs.gov	USGS
Schneider	Diane	diane_schneider@usgs.gov	USGS
Scholz	Donna	dscholz@usgs.gov	USGS
Schweitzer	Peter	pschweitzer@usgs.gov	USGS
Scott	Linda	lscott@usgs.gov	USGS

Last Name	First Name	E-mail	Affiliation
Sellers	Elizabeth	esellers@usgs.gov	USGS and NBII
Shin	Sharon	sharon_shin@usgs.gov	FGDC
Shirley	Jolene	jshirley@usgs.gov	USGS
Shivers	Steve	spshivers@usgs.gov	USGS
Simmons	Carol	carols@nrel.colostate.edu	Colo. State Univ.
Simpson	Annie	asimpson@usgs.gov	USGS
Skinner	Chris	cskinner@usgs.gov	USGS
Smith	Jonathan	jhsmith@usgs.gov	USGS
Smith	Steven	smsmith@usgs.gov	USGS
Snyder	Stephen	ssnyder@usgs.gov	USGS
Soderberg	Nancy	nsoderberg@usgs.gov	USGS
Soller	David	drsoller@usgs.gov	USGS
Sonenshein	Roy	sunshine@usgs.gov	USGS
Stamm	Nancy	nstamm@usgs.gov	USGS
Stapleton	Jo Anne	jastapleton@usgs.gov	USGS
Steele	Clint	csteele@usgs.gov	USGS
Stevens	Tyler	stevens@gcmd.nasa.gov	NASA
Stewart	Jana	jsstewar@usgs.gov	USGS
Stone	Sean	sstone@usgs.gov	USGS
Strobel	Michael	mstrobel@usgs.gov	USGS
Tepper	Dorothy	dtepper@usgs.gov	USGS
Tewalt	Susan	stewalt@usgs.gov	USGS
Thompson	Doug	cthompson1@usgs.gov	USGS
Thompson	Phyllis	pthompson@usgs.gov	USGS
Towns	Julia	jtowns@usgs.gov	USGS
Usery	E. Lynn	usery@usgs.gov	USGS
Vaughn	Alan	avaughn@usgs.gov	USGS
Wallace	Laure	lwallace@usgs.gov	USGS
Wenger	Etienne	etienne@ewenger.com	Learning for a Small Planet
Wertz	Robert	rwertz@usgs.gov	USGS
Wilson	Scott	scott_wilson@usgs.gov	USGS
Wimer	Mark	mwimer@usgs.gov	USGS
Wippich	Carol	cwippich@usgs.gov	USGS
Wood	Karen	kwood@usgs.gov	USGS
Wood	Kevin	wood@usgs.gov	USGS
Woosley	Lloyd	lwoosley@usgs.gov	USGS
Yoesting	Cheri	cyoesting@usgs.gov	USGS
Youman	Mark	myouman@icfconsulting.com	ICF Consulting
Zolly	Lisa	lisa_zolly@usgs.gov	USGS