## Executive Summary

GeoData 2011 was inspired by a joint NSF-USGS identification of the need to hear from the broader 'geo' community[1] on a variety of data related matters. While increasing attention needed to be paid to full life cycle of data, in the process of preparing and scoping the workshop two other hot issues were identified: integration and citation, giving the workshop three subject areas to delve into as well as to explore connections among them. Invited participants were drawn from all 'Geo' disciplines, and beyond, from information, computer and library science, from academia, agency and commercial organizations, and from student to senior faculty/ administrators. The workshop diversity provided a rich exchange of ideas, experiences and challenges for GeoData. Many key findings and recommendations have been extracted from the detail breakout discussions and syntheses during and after the workshop. Topical categories included: metadata, standards, standards-based tools, culture, collaboration and workforce.

Key points that cut across all three-subject areas were:

- A shift is needed within agencies to provide longer-term funding support, for communities to come together, remain coherent and to enable data stewardship, integration and citation within their communities and across to other communities (to the extent possible).
- Agencies like USGS, NASA and NOAA must also play a key role in sustaining geoscience cyberinfrastructure by moving research advances into operations.
- Community-wide standards and practices should build from demonstrated successes, be widely disseminated, and tools need to be developed to support them.
- Education is critical to broader adoption. Marketing studies need to be conducted to provide the business case for full stewardship, integration and citation, and incentives are needed to encourage everyone to participate in making data integratable, citable, etc.

While technology gaps are still evident across the three topic areas, there is recognition that human factors dominate, and often limit progress and effectiveness. Organizational and resource factors that would lead to a solid understanding of the business case behind, for example, making data preservable and integratable, come at a cost that is not well documented or supported when resource (funds and people) decisions are made.

The GeoData workshop exposed a strong commonality of challenges and potential solutions across agencies as well as with academia and the commercial sectors. This commonality makes it imperative that government agencies improve coordination amongst themselves and with the academic and commercial sectors in order to progress in the total life-cycle management and integration of science data.

In responding to recommendations herein, and as the GeoData community moves forward, deliberate attention must be directed toward cross-organization activities and discussions, including academic, governmental, and/or commercial. A scientific culture change in regard to modern data is underway but must be accelerated.

[1] Not to be confused with the GEO (Group on Earth Observations) community

## Introduction

Throughout the world, the management of data and information has emerged at the forefront of science strategies as organizations realize that investments have two important benefits: (1) entirely new types of scientific discovery become possible (Hey, Tansley and Tolle, 2010); (2) existing scientific practices become more efficient, allowing more science to be accomplished per dollar spent. This realization has resulted in many strategic plans, including "Harnessing the Power of Digital Data for Science and Society" (Interagency Working Group on Digital Data, 2009), "Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age" (National Academies, 2009), "Riding the wave: How Europe can gain from the rising tide of scientific data" (EU High Level Group on Digital Data, 2010) and "Interim Report of the International Council for Science's Strategic Coordinating Committee for Information and Data (SCCID, April 2011). Importantly, beyond attention to data management, research in data science is needed to provide the tools and knowledge to do this in a way that facilitates innovative science. In all of these plans, forums for data professionals to share experiences across a wide range of disciplines are recommended. These forums help build community, identify gaps and bright spots, and help define the path forward. In this spirit, the USGS and NSF Geosciences Directorate developed a workshop plan to bring together a diverse group of geodata providers, researchers and users to assess three aspects of the broad geoinformatics spectrum: the data life cycle, data citation practices, and data integration with strong emphasis on end-use.

*Data Life Cycle*: The data life cycle is a term coined to represent the entire process of data management. It starts with concept study and data collection, but importantly has no end, as data is continually repurposed, creating new data products that may be processed, distributed, discovered, analyzed and archived (see Figure 1). Fully supporting the different steps in the life cycle puts demands on metadata, standards, tools and people.

*Data Citation*: Data citation is the process of uniquely identifying datasets in a manner that can be indexed and sustained over time.

*Data Integration*: Data integration occurs when the Data Discovery and Data Analysis steps of the data life cycle integrate data from multiple sources, for example to pull data from a variety of distributed, heterogeneous sources to address complex issues such as climate.

## Workshop

The workshop was held March 2-4 in Broomfield, CO, with over 95 attendees from Government, Academia and the Private Sector invited by the Program Committee. Plenary sessions were interspersed with breakout groups on each of the three topic areas, and the workshop closed with a plenary panel discussion. The key findings and recommendations were developed by topic. Not surprisingly, all three topics raise issues for metadata, standards, standards-based tools, culture, collaboration and workforce.

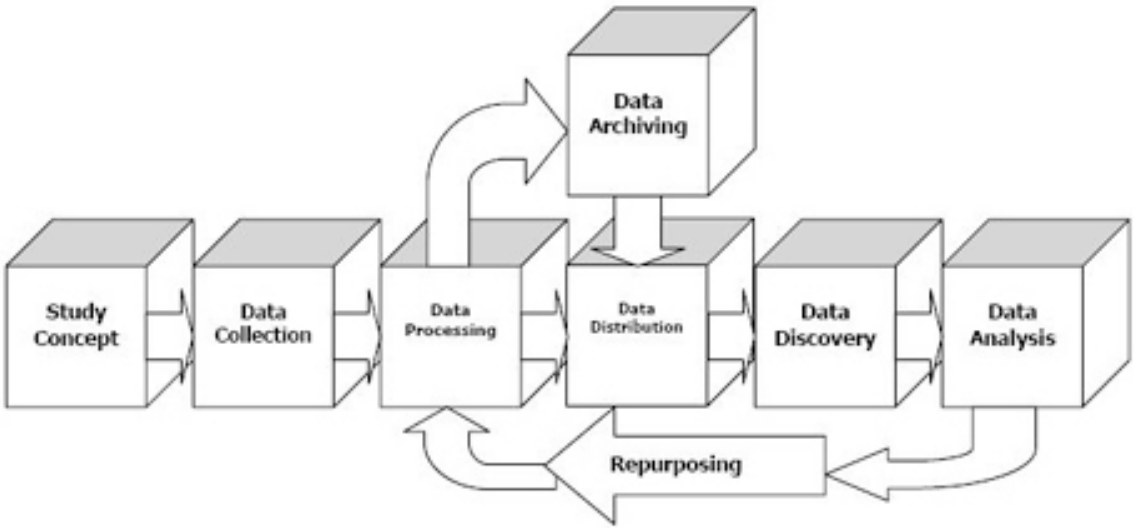


**Figure 1. The Data Life Cycle (figure from the MIT DDI Alliance).**

## Findings and Recommendations

### *Data Life Cycle*

*Key Life Cycle Stages*

- An important aspect of the data life is that it is not limited to data collection, archive, distribution/access and use. The data life cycle should begin with a planning phase at the front end, the "Study Concept" step in Fig. 1. Many workshop attendees advocated going further back to the proposal stage; a possibility with data management plans now being required for NSF proposals. To be effective at this stage, both proposers and reviewers must have enough knowledge about data management to fulfill their respective roles. This process can be greatly helped by ensuring that data management criteria are clearly set forth both in calls for proposals and charges to the reviewers.
- At the other end of the data life cycle, the aspects of use and reuse of data are understood up to a point, as indicated by the circular nature of the Digital Curation Center Life Cycle. The MIT DDI Alliance model takes this a step further, specifically calling out Repurposing the data as a feedback loop. These are both similar to the Spiral model used for software development. It is important to consider the implications of use and reuse. As a result, initial data collection planning focuses on the primary data users, i.e., those for whom the data are originally destined (often the data collectors themselves). Often, metadata to enable discovery and documentation to enable understanding are produced with this fairly narrow community in mind. Enabling secondary use by a wider community can be hampered by this narrow focus.
- The Repurposing step in the MIT DDI Alliance model implies a use of the data that was likely not envisioned at the beginning, or at least not in enough detail to plan for. As a result, design decisions made in metadata, data format, data structures and the like may be suboptimal for subsequent iterations of this feedback loop. Also, determining how much metadata and what type to add during these later iterations is not well understood in the

general data management community. In order to prepare as much as possible for secondary use, communities across domains will need to share anticipated requirements.

*Metadata at All Stages of the Data Life Cycle*

- Metadata plays a critical role at all stages of the life cycle, though the specific metadata issues may vary from one stage to the next. At the data creation stage, the emphasis is on collecting enough metadata. There is a balance between overwhelming the data collector (or provider) and obtaining too little metadata to get the job done. Indeed, some metadata, if missed at this stage, can never be recovered. One recommendation is to build metadata collection into the actual data collection apparatus, making the process as automated as possible. The metadata collected by digital cameras for the Exchangeable Image File Format (EXIF) was cited as an example of this. Establishment of community metadata standards and creation of applications and standard formats to facilitate collection are among some of the challenges.
- Another critical point in the data life cycle for metadata is the value-added processing that often occurs in the Reuse/Transform/Repurposing stage. There is little in the way of standards or even guidelines as to how much and what kind of metadata should be added at this stage. Suggestions ranged from providing narrative descriptions of the analyses, algorithms or other transformations used on the data as well as the importance of retaining original provenance.

*Best Practices in Data Life Cycle Management*

- One strategy that was noted by more than one breakout group is the identification of Best Practices in Data Life Cycle Management. This has potential payoffs for data management practitioners who can learn from each other's successes; for science researchers (data creators) who can learn about common expectations of their data and the needed metadata and documentation to aid data life cycle management; and for programmatic managers who can begin to establish benchmark practices based on collected Best Practices.
- A number of specific Best Practices were mentioned as examples during the workshop, such as the characteristics of the Thematic Real-Time Distributed Data Server (THREDDS) catalogs, whose simple hierarchical organization, human understandable names and dataset-specific access methods help data centers to work with their community of data providers. However, to be sustainable, a Best Practices repository is likely needed, with some kind of discovery (search and/or browse) available. To obtain submissions, a cross-disciplinary survey was recommended; awarding honors to "best" best practices may further encourage submissions.
- There were also a couple of key areas where Best Practices are either non-existent or not well known. For instance, a best practice in Data Preservation identifying clear guidelines on how to make decisions about what to keep would be helpful to the data management community.

*Collaboration and Communities of Practice*

- Besides Best Practices, another fruitful mechanism for improving data life cycle management is to leverage collaborations and Communities of Practice (CoP). Collaborations typically pair people from different disciplines such as domain scientists (with expertise in the specific science of the data in question) and informaticists, data managers or computer scientists. In the ideal case, the domain scientists provide key knowledge about the science domain and the use cases for the data, while the informaticists/data managers/computer scientists bring expertise in data management methods and technologies. Another suggested collaboration was between data management professionals and members of the library community; although there is some overlap in methods and technologies, the library community has a specific focus on some aspects of data curation and preservation (e.g., identifiers) that could be infused into the data management community. Likewise collaborations between academic scientists and their campus libraries on data life cycle issues are also useful.
- Communities of practice, on the other hand, have the effect of bringing together many people in the same discipline to share approaches and lessons learned (though these disciplines may be somewhat heterogeneous as well). One significant example of such a CoP is the Federation of Earth Science Information Partners (ESIP). However, there was some speculation that professional societies, such as the American Geophysical Union (through the Earth and Space Science Informatics Special Focus Group) might be able to play a similar role.

*The Human Factor in Data Life Cycle Management*

- Although it is tempting to seek technological solutions to the challenges facing us in data life cycle management, humans remain a key factor in the process. For instance, there was some discussion regarding the underuse of some standards. An example is the rich, but complex, series of ISO standards regarding metadata. This underuse may be partially due to the intimidating effect of the standards' complexity on prospective data providers. Tools to work with ISO metadata might help this, but are difficult to construct for the general case. On the other hand, domain-specific tools could lower the intimidation factor. (At the same time, excessive handwringing about the complexity of ISO standards might itself increase the intimidation factor.)
- Another aspect to consider is the tendency of people to continue working the way they have in the past: the inertia of habit. This was cited during the workshop for the case of slow growth in data submissions to archives. Both carrots (such as wider data use and data citations) and sticks (contingent funding and contingent publication of papers) were suggested as potential motivators to change this inertia.

*Training and Outreach*

- One of the key factors in addressing the Human Factor is a robust training and outreach effort. This is somewhat easier said than done: the target of such an effort is not a uniform group of scientists. It would be important to tailor the training or outreach to specific segment of the Geoscience data community, avoiding a one-size-fits all approach. One way

to do this might be via online videos and coursework similar to the extremely successful Kahn Academy.

- Informaticists, software engineers and data managers can benefit from more knowledge about data science, such as how data are used.
- Students should be taught early on the importance of proper methods of data collection and documentation: these represent our future scientific researchers. In addition, proper usage of data is also key. As in workforce training, the level of instruction and different emphasis will depend on the segment being addressed (e.g. Graduate students, Undergraduates or K-12)
- Citizen scientists and the public in general have similar needs to students in learning how to interpret scientific results and their basis in the data.

*Technology gaps*

- During the workshop, a number of primarily technological gaps were identified. Several of the gaps fell into the area of data provenance, which remains one where our reach exceeds our technological grasp. Mentioned earlier was the need for standards or guidelines on what metadata to add during the value-adding process. This is exacerbated by the challenge of identifying the level of granularity at which to document provenance: while the dataset level is too coarse on its own, documenting file level provenance is cumbersome when all upstream data are considered. In addition, the community has not yet come to final agreement on standards and guidelines for critical provenance components such as identifiers and locators (though there is some progress in this area within the ESIP, for example). As a result, establishing data "trust" through provenance is still a work in progress.
- Another general area that requires significant advances is in the area of cross-discipline discovery, which was identified as a possible Grand Challenge. This area is hampered by differences in terminology and metadata content, as well as some uncertainty in the actual use cases to drive this capability. In theory, semantic web technology should be able to help answer this challenge, but would require the development of ontologies that are more reusable across domains than the current inventory of ontologies.

*A business model for long-term archive persistence*

- One of the most critical gaps for data life cycle management that was identified during the workshop is the need for a Business Model for persistence of long-term archives. This was identified as problematic for the NSF model of funding infrastructure over the long term, and especially important given the requirement for Projects to have a data management plan.
- However, other agencies face similar challenges, as do federally funded research centers and consortia. Such a Business Model should take into account not just the funding of any supporting infrastructure over the long term, but also the attendant staffing, collaborations, data management research and development that are needed for data life cycle management to keep up with data and user growth, as well as technology evolution.

<u>Recommendations</u>:

- Develop review criteria for data management plans that recognize the criticality of upfront planning, design of metadata and data design, as well as the importance of enabling future repurposing of the data.
- Develop proposal and review criteria for instrument development that ensures adequate metadata collection at the time of measurement.
- Develop a searchable repository for best practices in data lifecycle management for capturing best practices noted in Geodata 2011, the NSF Research Data Lifecycle Management Workshop in July 2011, and related work in the Earth Science Information Partners and similar organizations.
- Identify Data Life Cycle Communities of Practice within NSF programs as well as other organizations such as ESIP and American Geophysical Union.
- Develop methods, such as workshops, AGU special sessions, etc. for bringing together overlapping Communities of Practice for information exchange.
- Develop domain-specific tools for adding and editing metadata, particularly within ISO metadata standards.
- Develop incentives (both carrots and sticks) to induce data providers to develop metadata and data products that will be usable by both narrow initial users of data and the wider community of interdisciplinary users reusing data for other purposes.
- Develop curricula targeted at both the practicing researcher and science students in data science, including data collection, management and integration.
- Continue supporting work in recording provenance of data, while shifting some resources to the practical application of this work.
- Support research and applications aimed at improving cross-disciplinary and inter-disciplinary discovery of data and related services.
- Develop Business Models for persistent long-term archives that take into account the funding cycle as well as the data life cycle.

## *Data Integration*

*Maturity of Data Integration*

- Between academic and agency realms, data integration and harmonization for sensors are mature and well documented, but for the "sample" or "lab" world are not. For example, data collected in the field by biological oceanographers.
- The strict notion of "interoperability" across all of science simply will not work but it can and is working within sectors of the science.
- Sustained investment in data integration and in supporting education and workforce training is needed – not just 2-3 year grant cycles.
- For many projects, two common themes emerged as being associated with some level of success in ability to do data integration: 'long-term' commitment of funding support AND active engagement of funding managers[2].
- There is a clear need to "Architect and Design" data pipelines toward integration, treating data like a supply chain problem.

[2] http://semanticommunity.info/A_NITRD_Dashboard/Designing_a_Digital_Future

- There is a need to build and manage to robust inventories not just metadata records.
- Data integration is a business problem, not an IT problem.
- Communities need to come together and communicate to achieve interoperability within their communities, and interoperability across them.
- Responsibilities for data integration are not well understood.
- To avoid information loss during data integration, data collectors and providers need future data integration enablers in their plans. Integration preparation steps need to be well documented (new metadata). Quality information needs to be carried along during the integration process (old metadata).

*The community and the agencies*

- The development of a data integration community of practice[3] is needed and must include
  - Infrastructure to foster communication (workshops)
  - Mentoring of students and early career PIs
  - Development of tools (e.g. Unidata developed NetCDF which has been adopted by many communities)
  - Education and training
  - The persistence and recognition of a 'named' community that can enable funds to flow from some agencies to researchers
- 'Long-term' funding support enables development of a community-of-practice that fosters communication, education and training, development and adoption of common tools and identification of core measurements. Communities-of-Practice can divide up the labor and work collaboratively to address shared challenges (economy of scale).

*Tools*

- A framework for understanding the use and production of open source software between academia and agencies would have positive downstream impacts on the dissemination and sharing of software to support data integration and the other themes of the workshop.
- The human factor must be addressed through tools if greater data integration is to be achieved.
- Metadata standards need to include mechanisms for integrating that information into tools.
- Tools need capability to capture lineage information automatically.

*Standards*

- Development and use of conventions and standards is still key and cuts across all the themes of the workshop. Incentives must be provided to encourage their development and use.
- SEED, netCDF, HDF and CF conventions are examples of working standards but gaps exist
- Interdisciplinary and workflow standards are needed
- Data integration would be better served if more metadata was captured at the point of data origin, as it is, for example, with EXIF for digital cameras

[3] Along the lines of the USGS Community for Data Integration (CDI), or Federation of Earth Science Information Partners (ESIP)

- Metadata needs to be evolved over time – it is not true that once created, it is forever good

*Semantics*

- Need to remember that integration means not only bringing together disparate sources of data, but also semantic integration as well
- Controlled vocabularies and ontologies are important

Recommendations:

- Sustain funding for discipline-oriented integration immediately. Agencies should consider data integration by discipline with other agencies and academic communities as a precursor to agency-wide data integration efforts. Encourage more participation in OGC Domain Working Groups.
- Develop a framework for understanding the use and production of open source software between academia and agencies. This would have positive downstream impact on the dissemination and sharing of software to support data integration and the other themes of the workshop.
- Develop a means to support the continued development of a community-of-practice that fosters communication, education and training, development and adoption of common tools and identification of core measurements.
- Enhance intern opportunities for students in agencies in Geo-Data informatics.
- Support translation/conversion/ontological tools that link communities.
- Expand the model of sustained funding for groups that build fundamental infrastructure for the community and are driven by community needs (Unidata model).
- Continue support for workshops like this and encourage continued partnership between agencies and academia.
- Expand research on how and where data management and curation approaches can apply for particular data types produced and used across disciplines. (Where we can implement general solutions, and where we shouldn't.)
- Fund social science studies in different geo-disciplines to more accurately document how much time scientists spend doing jobs that better geoinformatics could eliminate (browsing for data, reformatting data, writing custom codes to do routine operations). Question: How accurate are commonly cited estimates like 70/30 or 80/20 breakdown between data processing work and scientific analysis and understanding? Would likely estimate a price tag in the billions. How about "How to enable more science and make researchers more productive in an era of shrinking budget?" (This item is about marketing data integration).
- Foster traveling shows (seminars) to present what metadata can enable so that scientists can see the value. Community colleges could provide training on basic geoinformatics approaches, working together with masters level domain folks, interviewing/assisting scientists to enable quality metadata in their discipline. e.g., partnerships between community colleges and scientific institutions and internships at scientific institutions could be valuable. Try creative approaches: H&R Block helped Farmers supply geodata. Outsourcing, crowd sourcing, support work on dynamic, iterative approaches to metadata.

- Help foster academic programs in geoinformatics, sponsor summer internships (like Google Summer of Code, but for geoinformatics).
- Fund activities that get legacy data into modern interoperable formats and technologies

## *Data Citation*

*Diverse categories:*

There are many diverse categories of data that need to be citable. While many data collections, especially reference and community collections have well-established data management processes amenable to simple citation rules; other types of data are more problematic. Some examples from the workshop include:

- Biological organism classifications where an organism is in a classification scheme that changes as a function of time.
- Model runs for diverse application areas and needs.
- Seismic data organized by network, station, location, channel - granularity, continuous data versus events.
- Nimbus 2 satellite imagery on film - concluded that any digitized and cleaned up version of the data would be a new data set.
- Data and data subsets generated on the fly (e.g., reformatted, re-projected, visualizations, etc.).
- Aggregated data that is distributed by somebody other than the original archive
- Streaming data that is not permanently captured anywhere.

*Use dependent citation*

- General agreement was reached that citation content will probably be variable depending on the purpose for the citation. Reasons for citation discussed during the workshop included:
  - To give credit where credit is due,
  - To establish data authority,
  - For replication (verification of science results) purposes,
  - To aid in tracking the impact of the data set through reference in the scientific literature,
  - To advertise the existence of data, and
  - To facilitate data access.

  Of these reasons, in general it was agreed that citations that provide credit, facilitate impact tracking and data access are less demanding in terms of content than the other reasons to cite data and that for these purposes citing data sets should be encouraged.

*Sociotechnical challenges*

- Peer-reviewed publications (i.e. journals; commercial, professional society and others) have differing citation acceptance criteria and standards. These complicate citation implementation, e.g. would it be a reviewer or editor task to check data citations?

- Credit tracking in publications, journal impact factors and citation indices are a factor (including the emerging role of 'data journals') that needs further and detailed examination as the number of data citations is still very small and confined to sub-fields of 'geo'.
- Tools and standards are not always set up to handle citations (e.g., OGC services).
- There are incomplete mechanisms for making users aware of the citation and providing tools to generate or access a citation.
- What are the robust incentives for the data producers, and others?

*Unclear best practices*

- There is a lack of consensus on what exactly should be cited, what a citation means for data, and the granularity at which data should be cited.
- It is unclear what the role of aggregators (e.g. virtual observatories and other service providers) and distributors of data (e.g. community archives, national data centers, or even the World Data System) should be in data citation both from a technical and social perspective.
- Norms for generating citations for many of the problematic types of data mentioned above (e.g., on-the-fly data products, streaming data, data aggregates and composites etc.) need to be developed.

*Human factors and education*

- Data creators (e.g., PI's) need education on the subject of citation and must be included in the conversation, especially if they can persuade their colleagues to start to cite date but also since the citing of data follows social and professional norms. Good examples are needed urgently.
- Additional benefits to data creators are increasingly available by technical means, e.g. PI's would be able to track down exactly who used their data and how.
- With the increasing number of casual data users, reporters and such, a significant opportunity exists around using data citation to reach a popular audience to demonstrate the return of investment for the use of public funding for data generation.
- Education opportunities should be sought at all levels, i.e. not just universities but 2 and 4-year colleges, to middle and high school – get data in the classroom to take advantage of the present generation's increasing information and technical savvy – building in a sense that data are valuable and sources need to be cited.
- There is an urgent need to work with society publication committees and commercial publishers on data citation and it will be key to identify a credible group to advance this discussion (e.g. ESIP was one suggestion at the workshop).

<u>Recommendations</u>:

- Encourage community-based development of standard citation practices by making this a sub-topic for future community workshops.

- Journals should be encouraged to require that the source data for any published research must be archived and accessible; including ephemeral data sources such as streaming data, workflow results, and products generated on-the-fly.
- Encourage Communities of Practice, scientific unions such as AGU, the ESIP Federation and other such organizations to work with publishers to develop norms for data citation practice
- Encourage tool and standards organizations and communities to include support for data citations within their products.
- Encourage repositories and archives to provide mechanisms for generating and accessing citations for data as it is accessed.
- Identify key stakeholders for resolving the open questions and engage them, perhaps through a mechanism such as a workshop, in resolving these issues.
- Identify those groups leading education and training efforts for data, and advertise or develop their success stories, featuring the all-around benefits. Also identify gaps and potential solutions to address them.
- Distinguish whether different categories of citation have divergent or convergent forms by making this a sub-topic for a future community workshop.
- Seek data citation success stories from researchers, preferably across many geo-disciplines and publish them in science circles (e.g. AGU's EOS) and prominent web sites (e.g. the Polar Information Commons).
- Support the development of workshops, tutorials, and other methods of educating the research community in citing data.
- Ask ESIP to seek out a few early adopter, or eager society publication committees and commercial publishers with a view to opening up the dialog on data citation from a community viewpoint.
- Recommend that NSF require that Data Management plans submitted in proposals include explicit statements about the development of appropriate citations for any archived or published data.

## Summary of state of field and common gap identification

*Finding*: *Communities*

- A pattern of success seems to be fostering communities that work within a discipline, but across institutional boundaries, to define vocabularies and data models, and build useful tools (seismic community, met/ocean community, biological community). While one would like to believe that mutual benefit would keep these communities working together, the reality is that often progress stops when the funding dries up.

*Recommendation*: Sustain funding for discipline-oriented integration. Agencies should consider data integration by discipline with other agencies and academic communities instead of agency-wide data integration efforts. Encourage more participation in OGC Domain Working Groups.

*Finding*: *Cyberinfrastructure*

- While best progress is often made within communities, there are often cyberinfrastructure elements such as scientific feature types (images, grids, time series, point samples), concepts, technologies, and approaches that are common, and therefore building cross-walks between communities and enabling cross-fertilization is essential.

Recommendation: Support translation/conversion/ontological tools that link communities. Expand the model of sustained funding for groups that build fundamental infrastructure for the community and are driven by community needs (Unidata model). Continue support for workshops like this and encourage continued partnership between agencies and academia. Expand research on how and where data management and curation approaches can apply for particular data types produced and used across disciplines (Where we can implement general solutions, and where we shouldn't).

*Finding*: *Marketing*

- There is an open question of how to effectively market Geoinformatics? The usual focus is to state how geoinformatics will lead to 4th Paradigm breakthroughs (however, see Tim Killeen's challenge to the workshop earlier in this report): The open question was: could we also market geoinformatics on the potential to save $$$ or do more science by making "ordinary science" more efficient?

Recommendation: Fund social science studies in different geo-disciplines to more accurately document how much time scientists spend doing jobs that better geoinformatics could eliminate (browsing for data, reformating data, writing custom codes to do routine operations).

*Finding*: Where is funding allocated?

- It is often very hard to get scientists to provide metadata. Until scientists receive adequate credit for good and useful data (e.g. data citations count towards tenure, etc.) they will struggle to see the value and are too busy paying attention to the matters for which they are rewarded. A culture change will be required.

Recommendation:

- Traveling shows (seminars) to present what metadata can enable so that scientists can see the value.
- Community colleges could provide training on basic geoinformatics approaches, working together with masters level domain folks, interviewing/assisting scientists to enable quality metadata in their discipline. E.g., partnerships between community colleges and scientific institutions and internships at scientific institutions could be valuable.
- Try creative approaches: H&R Block helped Farmers supply geodata. Outsourcing, crowdsourcing, support work on dynamic, iterative approaches to metadata.

*Finding*: *People!*

- As geoinformatics grows, there is a very clear need for more proficient geoinformaticists to solve problems.

Recommendation:

- Help foster academic programs in geoinformatics, sponsor summer internships (like Google Summer of Code, but for geoinformatics).

**Overall summary of key points**:

- A shift is needed within agencies to provide longer-term funding support, for communities to come together, remain coherent and to enable integration within their communities and across to other communities (to the extent possible).
- Standards and practices should build from demonstrated successes, and tools need to be developed to support them.
- Education is critical to broader adoption, marketing studies need to be conducted to provide the business case for integration, and incentives are needed to encourage everyone to participate in making data integratable.

## Roles for the community and agencies

Moving forward on a comprehensive life cycle based solution for the challenges the "geo"sciences face will require an active and organized community approach. This and past workshops have advocated for the formation of a light community governance structure and the use of domain-based and theme-based working groups from across the community of practice to take on those high priority issues that need a common solution or set of solutions. This approach also applies to the need for a common community infrastructure for data life-cycle management. Sustainable archives of community data with services for search and integration are a high priority, as is an open and highly disseminated architecture that allows for institutions and individuals to join an infrastructure as a node for data and information while maintaining local control. Underlying standards and tools to support such an infrastructure were a central part of the dialogue at this workshop with education, metadata, and semantic solutions being dominant. The community will need to step forward to meet this challenge by allowing leadership to form, participate actively in working groups, and then to deploy agreed upon solutions. Because resources across institutions are not uniform, support will be needed to bring underserved domains and institutions to the table.

Federal agencies have a strong role as a fundamental part of their missions to deploy rigorous data lifecycle management in their agencies. Additionally they have a strong role in providing support to the community through funding, providing sustainable archives for community use, sharing infrastructure, and actively engaging in education and research that provides solutions. In addition to sustaining support for cyberinfrastructure through long-term funding, agencies like USGS and NOAA can also play a key role in sustaining geoscience cyberinfrastructure by moving research advances into operations. Coordination groups such as ESIP, and partnerships among agencies, academic institutions, and the private sector could facilitate the creation of the infrastructure and tools envisioned here. This would take a commitment of joint resources – people and funds – and a willingness to share and adopt data, tools, and research. By creating a shared vision of the priority problems to resolve and the goals to be reached within the next 10 years regarding data life-cycle management, the community could leverage its resources, find

solutions and establish accessibility to data quicker and more efficiently and possibly move science and innovation forward at a faster pace. Not only is the amount of data being generated overwhelming, but also the amount of data not available for responding to urgent societal needs is equally overwhelming and confounding. It is costing both our economy and well being to not have these data available. As a business practice, the amount of data and opportunities for science advancement that are lost because of the lack of a common infrastructure is debilitating. Interagency coordination at the highest level will ensure federal participation, but more importantly, a well-funded coordination office that works with the "working" level of agencies and academic institutions to facilitate working groups and workshops, adoption of standards and tools, and creation of sustainable archives could go a long way in providing the first step towards a shared vision.

Culturally this is a difficult task to take on. The community will need to overcome its predilection to not share ideas and resources and its desire to only use its own solutions. Incentives to share data and to work together will need to be developed to maintain and continuously improve a common infrastructure. Agencies and institutions will need to agree upon a different system of values in rewarding their staff and on valuing the preservation and dissemination of data and information.

*Recommendation:*

- Create an interagency working group to foster data life cycle management practices, support coordination and informatics science initiatives, and to develop a high level shared vision and strategy for data life cycle management
- Create and fund a coordination office that works with the "working" level of agencies and academic institutions to facilitate working groups and workshops, adoption of standards and tools, and creation of sustainable archives
- Support and actively participate in existing coordination groups, and create where needed, new communities of practice in data life cycle management across agencies, academic institutions, and the private sector
- Establish an NSF working group to address the issue of incentives and cultural change needed to facilitate implementation of data life-cycle management

## Suggested next steps

On the last morning of the workshop and especially leading into the closing panel summarization, the still packed room was looking forward. Agency representatives from NOAA, NASA, USGS and NSF were encouraging with new initiatives in their respective organizations. It is clear that there is a strong interest in partnering among agencies and partnering across the community. Current and emerging initiatives like NSF's Earth Cube, USGS's Community for Data Integration, and interagency initiatives in Coastal and Marine Spatial Planning are all providing platforms for partnering and potential for new research. There is a clear need to continue the multi-agency, diverse community participation activities, focused on maintaining the deliberate progress made at Geodata 2011. Agency sponsors and the community need to come together through initiative building and through facilitated dialogue such as workshops and town halls at

upcoming major science meetings to begin discussion of the numerous recommendations of this report and how to work together as a community to resolve the challenges we face. It is hoped that this report provides rich fodder for scientists to consider in building new initiatives in the coming years. Finally, the majority of meeting participants supported convening a GeoData 2012 to continue strengthening the community, examining progress, and discussing future priorities.

## Format of meeting, agenda, overview of talks, breakout questions, organizational structure

### *Format of meeting*

Prior to the workshop, all invitees were provided with background material on the Geo-Data life cycle, integration and citation. The "charge" to the speakers was be to prepare a ~ 40-45 mins talk on state-of-the art in their topic and how their topic would relate to the needs of geodata informatics.

The workshop was conducted with all participants (~100) in the same room for plenary sessions and partition into four facilitated breakout discussions of ~ 25 participants, and each focused on one specific question (chosen by organizers from Preparation and breakout questions, 1-4 below). The composition of the four breakout groups remained approximately the same for each of the breakout sessions. The first half day (PM) was organized around invited presentations (NSF keynote by Tim Killeen and data life cycle keynote by Peter Fox) and framing of initial goals, outcomes, and activities etc., followed by the first breakout session on Data Life Cycle. An evening welcome reception was held for participants to socialize, network and consider the tasks ahead. Day 2 commenced with reports from each of the day 1 breakouts. Then the plenary keynote on Data Integration by Jim Barrett (Enterprise Planning Solutions) was followed by an introduction to the breakout sessions, followed by four facilitated breakout discussions on data integration. In the afternoon, the third invited presentation by Mark Parsons (NSIDC) set the scene for the final detailed breakout discussions on data citation for the afternoon. Finally, on the last half-day (AM), both sets of day 2 breakouts reported their findings. Four agency perspectives (USGS, NASA, NOAA, and NSF) provided important reflections for the participants to consider as the workshop was coming to a close. These perspectives were both personal and with the specific agency influence. Finally, a closing plenary panel discussion featuring initial comments by Erin Robinson (ESIP), Mark Parsons (NSIDC), Rich Signell (USGS), Ted Haberman (NOAA/NGDC) and Chris Lynnes (NASA/GSFC) with questions/ answers and discussion session from participants.

### *Agenda (OMNI Hotel, Broomfield, CO, Mar 2-4, 2011)*

**March 2**

- noon-1pm lunch, meet and greet
- 1-1:05pm Introduction (Interlocken ballroom C/D)
- 1:05-1:40pm Opening talk, Dr. Tim Killeen (Assistant Director NSF/GEO)
- 1:40-2:25pm Plenary talk, data lifecycle, Peter Fox (RPI)

- 2:25-2:35pm Introduction to break out
- 2:35-5:30pm (with break) four breakouts on data lifecycle
- 6pm welcome reception

**March 3**

- 8:30-9:30am breakout review (data life cycle)
- 9:30-10:15am plenary talk, data integration, Jim Barrett (Enterprise Planning Solutions)
- 10:30-1:00pm four breakouts on data integration
- 1:00-2:00pm lunch
- 2:00-2:45pm plenary talk, data citation, Mark Parsons (National Snow and Ice Data Center)
- 2:45-5:45pm (with break) four breakouts on data citation
- dinner on your own (shuttles are available to get to nearby restaurants)

**March 4**

- 9am breakout review (data integration and data citation)
- ~ 10am agency perspective
- ~ 11am panel with breakout leads, agency reps. - initial workshop synthesis discussion, feedback, close of meeting
- noon-1pm lunch

Twitter: #geodata2011

Web site: http://tw.rpi.edu/web/Workshop/Community/GeoData2011 (includes links to all presentation materials, and working notes and summary presentations from each of the breakouts).

## *Preparation and breakout questions*

As part of the registration, people were asked to respond to the following four broad questions (input was used to determine breakout questions and guide discussions):

1. In your view/ experience what parts of data life cycle, data citation and data integration implementations/applications or frameworks are well established (or not) in your discipline(s) and what are the common gaps?

2. How would you give guidance or prioritize how to address gaps in the lifecycle of data acquisition; curation and preservation? For data citation and integration? Are there new program or community opportunities?

3. What do you see as the important elements of a communications strategy to meet the needs of the research and agency communities? Do these include demographically underserved communities (including aspects of management and data infrastructures)?

4. What opportunities or means are there for academic and agency collaboration in geoinformatics and Geo-Data informatics data life cycle, citation and integration?

How can they be optimally leveraged and implemented to serve the needs of all constituents?

Two weeks prior to the meeting the following email was sent to registered participants:

Review the Description, Outcomes and Deliverables for the meeting and consider the following questions (Life cycle, citation, integration):

- What is the one most important thing you need to address to make progress in your field/ job/ project?
- What is the one most important thing you are working on right now?
- What three things would you like to get out of the meeting?
- What three things do you expect to contribute to the meeting?

## *Organizational structure*

Co-Chairs
- Peter Fox* (Rensselaer Polytechnic Institute)
- Rich Signell* (USGS)

Science Organizing Committee:
- Ken Casey* (NOAA/NODC)
- Ruth Duerr* (NSIDC)
- Susan Carbotte (LDEO)
- Linda Gunderson* (USGS)
- John Helly (UCSD/SDSC)
- Kerstin Lehnert (LDEO)
- Fran Lightsom (USGS)
- Tom Loveland (EROS Data Center)
- Christopher Lynnes* (NASA/GSFC)
- Catherine Norton (MBL/WHOI library)
- Amy Stout (MIT library)
- Bruce Wilson (ORNL)
- Steven Worley (NCAR/CISL)

* Members of SOC executive.

## *Agency support and acknowledgements*

The co-chairs express appreciation to the science organizing committee for their contributions to this workshop and the SOC executive (noted below)for assembling the workshop report. The co-chairs and science organizing committee for this workshop wish to express our significant appreciation to the NSF as primary financial sponsor, and USGS as supporting sponsor as well as all participants. In particular we thank:

NSF leadership: Eva Zanzerkia, Leonard Johnson, Cliff Jacobs, Robert Dietrick, Tim Killeen
USGS leadership: Linda Gundersen, Kevin Gallagher
Administration (RPI): Jacky Carley, Pamela Muraka, Carol Trifaro, Colleen Martin
Administration (USGS): Cheryl A Davis, Jennifer L Roberts, Lisa Ann Jordan,
Technical (RPI): Stephan Zednik, Eric Rozell, Patrick West

Technical (USGS): Timothy H Lee, Thomas W Van Dreser
LOC (RPI): Joanne Luciano, Stephan Zednik
Venue and Audio-visual/ network: Ashley Vap, Tamara Full (OMNI Interlocken), Colin Burke and staff (AVT)

## References

The Fourth Paradigm: Data Intensive Scientific Discovery, Eds. Tony Hey, Stewart Tansley and Kristin Tolle, Microsoft External Research (2009) - http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Harnessing the Power of Digital Data for Science and Society", Interagency Working Group on Digital Data (2009) - http://www.nitrd.gov/About/Harnessing_Power_Web.pdf

Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age, National Academies Press (2009) - http://www.nap.edu/catalog.php?record_id=12615

Riding the wave: How Europe can gain from the rising tide of scientific data", European Commission High Level Group on Digital Data (2010) - http://ec.europa.eu/information_society/newsroom/cf/document.cfm?action=display&doc_id=707

## Attendees

Participants: http://tw.rpi.edu/web/Workshop/Community/GeoData2011/Participants