MICROSOFT CORPORATION RESPONSE TO OSTP REQUEST FOR INFORMATION: PUBLIC ACCESS TO DIGITAL DATA RESULTING FROM FEDERALLY FUNDED SCIENTIFIC RESEARCH (FR DOC. 2011-28621)

SUBMITTED JANUARY 12, 2012

SUMMARY

Microsoft believes that curating, preserving, and using the digital data that result from federally funded scientific research are critical for advances in scientific discovery and for building a strong, innovative economy. We support the good work done to date by the research community and Federal agencies to define the challenges and outline possible solutions. In particular, we cite the report of the National Science Foundation's Advisory Committee for Cyberinfrastructure's Task Force on Data and Visualization (http://www.nsf.gov/od/oci/taskforces/TaskForceReport_Data.pdf) and the Blue Ribbon Task Force on Sustainable Digital Preservation and Access (http://brtf.sdsc.edu/). Experts from Microsoft participated in the drafting of these reports, and we remain committed to their conclusions. We also agree with many of the challenges described and conclusions reached in the National Science Board's draft Data Policies Report released on January 5, 2012.

As the National Science and Technology Council's (NSTC's) Interagency Working Group on Digital Data proceeds with deliberations to inform Federal policies concerning access to digital data resulting from federally funded scientific research, Microsoft would like to draw your attention to two areas—**Economic Models** and **Software Tools and Online Services**—and offer three recommendations in these areas.

Economic Models. The nation must create an environment in which innovation can occur around the critical elements that enable data sharing, retention and use, and costs can be distributed among the various groups that receive benefits from the data and associated discoveries. Challenges include the long-term nature of the problem (costs and activities in this space will extend over timelines greater than a typical research grant), and the need to make choices around what data should be preserved and shared. A wide variety of groups will create and use the data and they all must share in the costs and decisions about how data is preserved and shared. These participants include scientific communities and research groups of various sizes, universities, Federal laboratories, commercial service providers, and both Federal and state governments. They also include the consumers of the data, who may be outside of the research community, but who will have a stake in defining what data is of value and a responsibility to contribute to costs.

The economic models deployed around the data ecosystem will vary by discipline, but approaches should incent sharing and should provide support for the individuals and organizations that create, make available, and maintain high value scientific data collections. Exploration of different business models is critical, and the full variety of information technology (IT) infrastructures available for the various stages of data preservation and use should be exploited.

Recommendation 1 – Assessment of Economic Models around Data Retention and Sharing: Assessment projects should be undertaken to evaluate the *economics* associated with supporting and facilitating the long-term hosting and use of data. These projects should analyze potential economic models, including factors such as cost effectiveness, opportunities and risks for businesses and research institutions, and potential for value-added software tools and services. The results of these analyses should provide options for policies and programs through which the Federal Government might successfully foster a stable long-term ecosystem of service and data providers and consumers.

Recommendation 2 – Broaden the Range of Supported Information Technology Infrastructures around Data-Related Activities: Research communities and institutions, as well as information technology service providers, would be better able to explore different models for data sharing if there was clarification of Federal policies for support of information technology infrastructure, including computing services such as cloud, around data-related activities during and after research grants. In particular, it is important that Federal policies focus on the desired outcome (e.g. data sharing to advance science) and enable a variety of approaches to accessing the necessary IT hardware and software capabilities, whether through a purchase as part of a research grant, as an ongoing service from a commercial provider, or as an institutional or community resource.

Software Tools and Online Services. Simple, easy-to-use software tools and online services for data archiving, dissemination, discovery, and analysis are critical to maximizing the ability of researchers, entrepreneurs, companies and others to extract understanding and value from data. Also, metadata standards are key to facilitating the development and use of broadly applicable tools. Such tools are particularly critical for science conducted in increasingly multi-disciplinary and international environments. Some scientific disciplines have made progress in this regard, but more attention to this problem is needed.

Recommendation 3 – Support for the Development and Sharing of Software Tools and Online Services: Financial support is needed for the purchase, development and deployment of software tools and online services for data archiving, dissemination, discovery, and analysis. This support may take the form of Federal or foundation grants to universities or domain research collaborations, and such investments should include tools and services that can be shared across and customized for multiple research communities.

This issue relates back to Recommendation 1, as the role of tools and services is critical in evaluating potential economic models for data sharing. For example, a tool or service infrastructure that enhances the value of the data may allow the provider to monetize access to the data at a level sufficient to cover the investment made in creating or maintaining the data archive.

Additional Issues. In addition to the specific issues and recommendations discussed above, there are a number of other important policies that could contribute to supporting the effective long-term stewardship of publicly-funded research data. Examples include incentives to change the scientific culture around hoarding of data; development and implementation of domain-specific metadata, standards, formats, and protocols; rules around timing for and constraints to access to data by other researchers; ways to allow citations to data and credit to data creators and sharers; clarification of research agency expectations around domain-specific retention policies; and assigning responsibilities for long-term data curation. Although we do not discuss these topics in this document, we support the analysis and comments on them in the existing reports referenced above.

ECONOMIC MODELS

To maximize the opportunity for scientific discovery and innovation, it is favorable for data to be accessible and usable by a range of stakeholders including academic researchers, industry laboratories, members of the general public and international research collaborations. Sharing data advances scientific discovery, but also has positive economic impact, informs policy formulation, and provides educational and other societal benefits.

The growing amount of data created and used in scientific research is well documented, and the value and impact of making this data available has been widely discussed.[1] However, where the data (including metadata, images,

[1] For examples of how using new computing capabilities to explore massive datasets will advance multiple scientific domains, see *The Fourth Paradigm: Data-Intensive Scientific Discovery* (http://research.microsoft.com/en-us/collaboration/fourthparadigm/default.aspx).

and software tools) will be stored, how the use of this information will be facilitated, and who will pay to develop, deploy, maintain, and improve the relevant resources and capabilities remain a fundamental challenge. It is unlikely that a single model will work across all disciplines, research team sizes, institutions, and countries.

All Data are Not Equal

It is not affordable, or even possible, to save all data from all scientific investigations for all time. It is important to understand when data has re-use potential and when it does not, when it is easier or cheaper to recreate data than to store it, and when data that has been saved is no longer worth keeping. A variety of factors will inform such decisions, including how the costs of storage, access, and use are paid. The choices and the associated costs are not obvious – cheaper storage media, with limited bandwidth access might be an option for infrequently used information, while more intensely used information can need a variety of additional services, such as low latency storage system, replicated storage at multiple sites, high bandwidth connections, and tools and computing power to search, analyze, and access the data.

Another key variable in data retention and sharing discussions is the source of the data. Not only will different scientific fields have different cultures, priorities, and expectations around data management, but different research models also will require different approaches. For example, for data generated by single investigators or small groups, the processes and infrastructure for sharing may be a significant burden on the researcher. In this case policies and tools could focus on methods that allow the research to move the data into an existing curated collection that has a well sustained business model. For larger, multi-disciplinary or multi-institution collaborations, some level of shared data storage, access, and analysis is likely to exist as part of the collaborative process and sustainability rests on the economics of the discipline. For example, high energy physics and astronomy have models in which there is long-range government funding. In other areas, the users of the data may come from outside the scientific community generating the data – from other research fields or from commercial entities. Biology and chemistry have both nationally supported archives as well as the potential for public-private partnerships. In these cases, decisions about what data to keep, how to disseminate data and associated services, and what parties bear which costs requires negotiations beyond the community of the original data creators.

International partnerships are a special case. Different countries have different cultural norms, different policy mandates, and different economic models around the responsibilities of governments, universities, and other organizations. Flexibility will be necessary to craft solutions that balance different requirements, and up-front planning for the systems, costs, and policies of data retention and access will be critical.

Enabling Different Models – Facilitating Public-Private Partnerships

In addition to being cognizant of the variety of services and support that could be associated with a given data collection, it is important to recognize the goals, resources and priorities of the individuals, communities, or institutions that are either producing the data, using the data, re-using the data, storing the data, providing tools around the data, or paying for these various activities. Consequently it is favorable to enable economic models that have the flexibility to allow different groups to provide different services to different audiences at different times and costs.

There are a number of models that provide support at different stages of the data lifecycle (see below), but allowing market forces to operate can be of help in understanding how to preserve data and what kind of access patterns need to be supported. In particular, flexible pricing models can be important for gathering quantitative information about which data sets are being used, by whom, and how. Gathering such information facilitates informed evolution of choices about retention and pricing and could be used, in concert with evaluation of scientific needs and directions, to determine when data should be moved to cold storage or expunged.

A variety of approaches to providing and maintaining data ecosystems can be imagined or observed. For example, Federal agencies can create and operate systems for storage and access, or pay third parties directly for those services. Research organizations, including universities or scientific societies, can provide organization-wide access to services funded by fees or supported out of indirect costs. Individual researchers can fund data dissemination from a specific research grant, and can pay for access to data or tools on an ad hoc basis. Examples of approaches already deployed in various fields include:

- The University of Michigan Inter-university consortium of political and social research, asks customer institutions, such as universities and research laboratories to subscribe on behalf of their researchers.
- The data from the experiments on the Large Hadron Collider is managed by an international, multi-tier distribution system which is funded as part of the project and is provided for free to the participating physicists.
- For a number of geosciences and life sciences data sets, there is already a marketplace for providing access to data based on modest subscription fees which cover the cost of maintaining high quality tools to search, analyze and access the data as well as storage costs for the data. Examples include Datamarket.com and Windows Azure DataMarket, LifeSpan BioSciences, Inc.

No specific model is correct for all situations, but the most important factor in ensuring successful data impact is enabling various organizations to bring their skills and resources to different elements of the data lifecycle. Public-private partnerships will be an important component. For example, in some situations, federal agencies may directly fund the research that generates data, but later only indirectly fund the storage and access to that data by allowing other researchers to pay fees for access and tools developed and maintained by the original researchers, other researchers, a scientific society, or a for-profit entity. An additional market value is created when organizations can develop and deploy value-added services on top of free data, or data from other organizations.[2]

The Role of Cloud Computing and Storage

Cloud technologies, which are being developed and deployed for a variety of business, government, and consumer applications, are relevant to the data challenges in a variety of ways.

Move the Analysis to the Data: Today, a scientist can store or download modest amounts of data to a local computer for limited analysis and study, but increasingly the size of the data sets or the computing power required for analysis will make this inefficient or impossible. It will become necessary to move the analyses to where the data is. Using the large data centers that have been built to support massively parallel analysis of resident data, scientists can conduct research on petascale data archives in ways that are not possible on local facilities.

Environment for Collaboration: Cloud computing services potentially provide an information technology environment that facilitates both collaboration and effective data sharing. This may be particularly valuable for multidisciplinary and/or multi-organization collaborations. Cloud computing may also be a platform to support the ecosystem of data sharing and use – different parties can come together in the cloud to provide different elements of the tools and services needed (from the data, to the storage, to the applications and tools, to the computational power). Microsoft maintains a cloud based marketplace for data access. While some of the data is subscription based, there is a great deal of public data that is provided free of charge. Other large cloud providers, such as Amazon and Google, offer similar services.

Build on Other Investments: The scale of cloud deployments, and the rapidly evolving ecosystem around cloud applications for business, government, and consumers, as well as science, are likely to facilitate the evolving

[2] The European Union emphasized the potential economic value of commercial re-use of data in announcing its proposed Open Data Strategy in December (http://www.zdnet.co.uk/news/regulation/2011/12/12/reuse-of-public-data-to-get-easier-under-new-eu-rules-40094628/).

development and deployment of technology that can potentially reduce the costs of data storage, access, and analysis for research.

Recommendations (Economic Models)

An important step in creating a vibrant environment for data sharing is facilitating people and organizations' ability to experiment with different approaches for different research communities. In particular, it is critical to enable *an ecosystem in which different actors can contribute relevant materials, tools, and products for the different elements of the data lifecycle*.

Recommendation 1: Assessment of Economic Models around Data Retention and Sharing
To encourage the development of an ecosystem of services supporting data retention, access, and use, Federal agencies should support targeted economic assessment projects. The goal of the projects would be to explore the economic viability of a variety of support models for the longer-term hosting and use of scientific data including both academic use and any potential commercial exploitation that could be used to supplement (or completely pay for) the costs the data access and retention for academic research. The results of these analyses should provide options for policies and programs through which the Federal Government might successfully foster a stable long-term ecosystem of service and data providers and consumers.

Types of Projects: The assessments should explore a variety of disciplines and consider the roles and needs of single or small group investigators, multi-disciplinary/multi-institution or public-private collaborations, and data collected for scientific and operational consumption across a variety of sectors. For example, one assessment might explore the use of the cloud by a multi-institution collaboration for its own data analysis in the short-term as well as dissemination of data and associated tools to the larger community in the long-term. Another might look at the development and deployment of tools that cost-effectively allow single investigators to manage the data lifecycle and workflow from creation to archival. Another might evaluate the business models for how weather data is used by climate researchers, government operations, and commercial entities.

Questions to Be Explored:
- how users as well as the disseminators of data are currently supporting the associated costs of data management – to give a clear understanding of the current cost 'baseline';
- tracking of who is using shared data and for what purposes (including access patterns and derived value), as well as how the users discovered the data – to understand current practices and successful collaborative models;
- existing or potential inflexion points in data management costs due to economies of scale (e.g. data or access volumes which suggest a more cost-effective transition to cloud-based service providers) – to understand the criteria for when/where different types of service are applicable for different volumes of data or where access volumes might experience different service levels/support costs;
- cost-effectiveness of different service models and comparison between cost structures across the different phases of the data management life-cycle – to understand whether/where there are particular models which work for specific parts of the data access/use/retrieval lifecycle and how these differ between scientific disciplines;
- potential for value-added software tools and/or services – to understand what scope there is for software or service infrastructure might be applicable/available to different types of research data and, specifically, whether there would be opportunities for commercial exploitation of data which could supplement or completely cover the costs of data retention and access by the research community;

The gathered data and assessments could inform Federal decisions on what sorts of programs would enable data ecosystems. In addition, the domain-specific information could help inform research groups considering long-term data management plans of various relevant business and information technology models.

Recommendation 2: Broaden the Range of Supported Information Technology Infrastructures around Data-Related Activities.

The fiscal role of the Federal government in enabling access to data can take many forms, including direct and indirect support of various elements of the data sharing ecosystems. Research communities and institutions, as well as information technology service providers, would be better able to explore different models for data sharing if there was clarification of Federal policies for support of information technology infrastructure, around data-related activities during and after research grants.

In particular, it is important that Federal policies focus on the desired outcome (e.g. data sharing to advance science) and be flexible about which IT infrastructures are used to achieve the outcome. Conducting research, creating data, preserving, sharing and reusing that data can require a wide array of IT hardware and software capabilities, and these capabilities can be achieved in a variety of ways—purchased through an individual research grant, provided by a university as an institutional resource, obtained through a community resource shared across a scientific domain, or acquired on an as needed basis from a commercial service provider. When architecting any new Federal policies, it would be advantageous to avoid discriminating against any particular approach and/or presuming a favored solution.

For example, in determining what IT infrastructure is allowed to be used in the conduct of a research grant, a selection should take into account not only the resources necessary to carry out the specific research, but also whether the choice will smooth the transition for data to be shared. This could include support within the grant for usage (and fees) for community resources, or payment for commercial storage, computing, or software services.[3]

Software Tools and Online Services

The development and deployment of software tools and services to enable sharing, discovery, and analysis of data is key to realizing the ultimate goal – increased scientific and societal impact of data. Unfortunately, beyond a few tools used within a few narrow scientific subdomains, there are no standard software packages that scientists can use today. Metadata standards are equally scarce. Researchers, government agencies, companies, and others have a role to play in creating and supporting these tools. The deployment of on-line services that provide these essential capabilities increases the value of the data and creates a possible market and business model to sustain it.

Software tools are critical at every stage of the data sharing lifecycle. From the start, tools that simplify the steps of preparing data and associated metadata for sharing, perhaps integrated into the data generation and capture process, facilitate data preservation, annotation, and sharing. If the data workflow is well managed and cataloged from creation, then archival requires much less effort.

Increasingly, the value of data extends beyond traditional disciplinary boundaries, and ensuring data access and the ability to correlate data from multiple disciplines requires appropriate metadata, protocols, and interoperability standards. Tools for analysis that are deployed in concert and coordination with specific data sets are particularly valuable in supplementing metadata and other annotation of data sets. Specific technical issues that must be addressed include robust, long-term secure digital storage, reliable techniques for predicting storage media aging, mining large-scale data collections to provide useful information, and visualization tools.

---

[3] In particular, purchase of IT equipment and purchase of IT services may be treated differently under Federal grant regulations (OMB Circular A-21) in terms of whether they will be subject to indirect costs, providing a potential fiscal disincentive to utilize services. However, in certain circumstances the use of services, such as cloud computing, may be a more effective approach to meeting goals for timeliness of access to resources, flexibility in scaling storage or computing power up and down, enabling of long-term data retention, etc.

In addition to coupling data with analysis tools, great value can also be obtained by connecting data with research publications. As these publications become digital artifacts, it should become easier to trace the provenance of a research result back to the supporting data collections and analysis tools. This capability will facilitate repeatability and reproducibility in scientific experiments and transparency around the context when data and analyses are being used in policy discussions.

Recommendation 3: Support for the Development and Sharing of Software Tools and Online Services

Software tools are essential for facilitating data archiving, dissemination, discovery, and analysis. Federal programs and policies should facilitate access to and use of such tools and online services. This should include financial support for the purchase, development and deployment of software tools and associated online services, which may take the form of Federal or foundation grants to universities or domain research collaborations. It also should include policies designed to minimize duplication of existing resources and encourage the sharing and reuse of tools and services.

- *Use of Existing Capabilities*: Federal research programs could give priority to research proposals that describe how they will make use of commercially available data services and software tools or community developed and supported tools in cases where these can cost-effectively provide storage, analysis capabilities, visualization, pre/post processing and/or data handling/manipulation capabilities.
- *Focus on Sharing*: Priority for Federal investments could focus on tools and services that can be shared across and customized for multiple research communities and domains of science. Funding tool development separately from domain science would encourage a focus on capabilities relevant to multiple fields.
- *Stable Long-Term Deployment and Use*: Proposals for tool development support should include consideration of the potential deployment models, including the opportunity for a sustainable economic model for the maintenance of the tools. Emphasis should also be placed up front on what metadata will be required for effective use of the tools.

Examples of the types of tools and services that are important for researchers in multiple fields to have access to include tools that simplify the data lifecycle management including the steps of preparing data and associated metadata for sharing, easy-to-use search, visualization, and analysis tools, and tools for that allow limited access and analysis of data to maintain privacy preservation[4] or intellectual property constraints.

[4] A potentially significant barrier to data sharing in certain research areas, such as biomedicine and some social sciences, is concerns about maintaining compliance with related privacy and data integrity rules. Examples of the challenges include the potential for de-anonymization of data when multiple related data sets are shared, or the need to comply with different country-specific regulations. Tools to constrain users to acceptable analyses, or methods to build data sharing approaches around providing analytical results rather than raw data, are needed to facilitate the realization of the economic and societal benefits of data sharing and reuse.