Chapter Three: Roles and Responsibilities of Individuals and Institutions

SHARED GOALS AND RESPONSIBILITIES

Sound policy development and implementation rest on the recognition of the roles and responsibilities of those who play an active part in the digital data collection universe – the data users, authors, managers, and funding agencies. One of the goals of policy is to ensure that these roles and responsibilities are clearly defined and properly fulfilled. In pursuing their respective interests in data collections, each actor in the data collection universe has a distinct set of responsibilities, which are outlined in the paragraphs that follow. In addition to their separate responsibilities, the groups must also act collectively to pursue some of the higher-level goals important to the entire fields. Examples of such goals are the following:

- ensure that all legal obligations and community expectations for protecting privacy, security, and intellectual property are fully met;
- participate in the development of community standards for data collection, deposition, use, maintenance, and migration;
- work towards interoperability between communities and encourage cross-disciplinary data integration;
- ensure that community decisions about data collections take into account the needs of users outside the community;
- encourage free and open access wherever feasible; and
- provide incentives, rewards, and recognition for scientists who share and archive data.

An important policy consideration is the creation of opportunities and mechanisms by which all of the groups can work together in addressing universal goals.

DATA AUTHORS

The interests of the data authors – the scientists, educators, students, and others involved in research that produces digital data – lie in ensuring that they enjoy the benefits of their own work, including gaining appropriate credit and recognition, and that their results can be broadly disseminated and safely archived. In pursuing these interests, the data authors have the following responsibilities:

- conform to community standards for recording data and metadata that adequately describe the context and quality of the data and help others find and use the data;
- allow free and open access to data consistent with accepted standards for proper attribution and credit, subject to fair opportunity to exploit the results

of one's own research and appropriate legal standards for protecting security, privacy and intellectual property rights;
- conform to community standards for the type, quality, and content of data, including associated metadata, for deposition in relevant data collections;
- meet the requirements for data management specified in grants, contracts, and cooperative agreements with funding agencies; and
- develop and continuously refine a data management plan that describes the intended duration and migration path of the data.

Robust, comprehensive, and broadly endorsed and disseminated community standards are crucial to the ability of authors to meet these responsibilities. Thus, active support for the development of community standards is an important policy goal.

DATA MANAGERS

Data managers – the organizations and data scientists responsible for database operation and maintenance – have the responsibility to:
- be a reliable and competent partner in data archiving and preservation, while maintaining open and effective communication with the served community;
- participate in the development of community standards including format, content (including metadata), and quality assessment and control;
- ensure that the community standards referenced above are universally applied to data submissions and that updated standards are reflected back into the data in a timely way;
- provide for the integrity, reliability, and preservation of the collection by developing and implementing plans for backup, migration, maintenance, and all aspects of change control;
- implement community standards through processes such as curation, annotation, technical standards development and implementation, quality analysis, and peer-review (some of these functions, defined in this report as community-proxy functions, apply primarily to resource and reference collections and may not apply to many research collections);
- provide for the security of the collection;
- provide mechanisms for limiting access to protect property rights, confidentiality, privacy, and to enable other restrictions as necessary or appropriate;
- encourage data deposition by authors by making it as easy as possible to submit data; and
- provide appropriate contextual information including cross-references to other data sources.

To be successful, the data manager must gain the trust of the community that the collection serves. Thus, collections policy should emphasize the role of the community in working with data managers.

DATA SCIENTISTS

The interests of data scientists – the information and computer scientists, database and software engineers and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection – lie in having their creativity and intellectual contributions fully recognized. In pursuing these interests, they have the responsibility to:

- conduct creative inquiry and analysis;
- enhance through consultation, collaboration, and coordination the ability of others to conduct research and education using digital data collections;
- be at the forefront in developing innovative concepts in database technology and information sciences, including methods for data visualization and information discovery, and applying these in the fields of science and education relevant to the collection;
- implement best practices and technology;
- serve as a mentor to beginning or transitioning investigators, students and others interested in pursuing data science; and
- design and implement education and outreach programs that make the benefits of data collections and digital information science available to the broadest possible range of researchers, educators, students, and the general public.

Almost all long-lived digital data collections contain data that are materially different: text, electro-optical images, x-ray images, spatial coordinates, topographical maps, acoustic returns, and hyper-spectral images. In some cases, it has been the data scientist who has determined how to register one category of representation against another and how to cross-check and combine the metadata to ensure accurate feature registration. Likewise, there have been cases of data scientists developing a model that permits representation of behavior at very different levels to be integrated. Research insights can arise from the deep understanding of the data scientist of the fundamental nature of the representation. Such insights complement the insights of the domain expert. As a result, data scientists sometimes are primary contributors to research progress. Their contribution should be documented and recognized. One means for recognition is through publication, i.e., refereed papers in which they are among the leading authors.

DATA USERS

The interests of data users – construed here to include the larger scientific and education communities, including their representative professional and scientific communities – lie in having ready access to data sets that are searchable, robust, well defined, and well documented. In pursuing these interests, data users have the responsibility to:

- adhere to appropriate standards for attribution and credit in the use of data generated by others and observe appropriate limits on redistribution;
- report significant errors to data managers or authors as appropriate;
- provide primary input to decisions on what data are valuable to archive (for instance, raw versus processed data) and for how long;
- reach consensus on data center needs/structure for their user community and evaluate the quality of the available centers; and
- respect restrictions on use, such as copyright and no-derivatives, placed on data sets.

Meeting responsibilities for attribution and for respecting restrictions on use requires that the relevant information be readily available to the user. Thus, an important policy consideration is the development of metadata systems that provide authorship, versioning, modification, licensing, and other relevant information. The system of digital licensing being developed by Creative Commons (see http://www.creativecommons.org) provides an example in this regard.

FUNDING AGENCIES

Much of the data currently being collected are ‘born-digital’ and lack any analog counterpart. Additional data are being converted to digital form and, in the process, are often dissociated from their analog representation. The digital data, and the investments made in gathering them, could be lost unless a robust preservation plan is created for digital data. This is the role and responsibility of NSF and other funding agencies, working in concert with data authors, managers, and users to:

- create a culture in which digital data receives the same consideration as data published in print form so that an author’s contribution is judged by the insights, creativity, and significance of the analysis and not by the media in which the data are created and stored, [compiling, editing, and publishing data in a data collection should be seen as a fundamental research responsibility. The emphasis on preservation (and the development of a stable preservation infrastructure) would be the equivalent to that now attached to the preservation of data in printed form];
- catalyze the creation of an accessible digital commons for research and education that provides the foundation for launching, operating, and preserving research, resource, and reference collections;

- support interactions within and between communities to allow the development of robust community standards for digital data and interoperability and facilitate the development of community norms, customs, and expectations for digital research; and
- enable the broadest possible access to the digital research environment by ensuring that both the physical resources and the necessary training are broadly available; provide the oversight to ensure that this training supports the development of the expert workforce and scientific leadership required for innovative digital discovery through digital data systems and collections.

The Foundation is in a unique position to act because of the fundamental support it provides for the research and education enterprise, its history of leadership in the area of digital data and research, and the breadth of disciplinary representation and participation found across the Foundation. Because digital data collections have become indispensable to advances in research and education, the task force believes that urgent action, involving transformative, rather than incremental, change is required.

DATA QUALITY ACT

Federal agencies have responsibilities under the so-called 'Data Quality Act' (Public Law 106-554; H.R. 5658, Sec. 515). In accordance with the Act, the Office of Management and Budget (OMB) has issued guidelines that "provide policy and procedural guidance to Federal agencies for ensuring and maximizing the quality, objectivity, utility, and integrity of information ... disseminated by Federal agencies" (see http://www.whitehouse.gov/omb/fedreg/reproducible2.pdf). These guidelines apply to information whose collection and dissemination to the public is initiated or sponsored by a Federal agency. NSF examples include the biennial *Science and Engineering Indicators* report and certain other publications produced by the NSF Division of Science and Engineering Statistics.

Importantly, the OMB guidelines do not apply to information disseminated by a Federal grantee or contractor or Federally employed scientist when he or she publishes and communicates research findings in the same manner as academic colleagues, or decides whether to disseminate research results or other data and what information will be included in the dissemination. Thus, the guidelines do not apply to information disseminated by NSF-funded grantees as outlined in the NSF Information Quality Guidelines (see http://www.nsf.gov/policies/nsfinfoqual.pdf):

> *NSF grantees are wholly responsible for conducting their project activities and preparing the results for publication or other distribution. NSF promotes data sharing by its grantees through its data sharing policy and by data archiving by its grantees. NSF does not create, endorse, or approve such data or research materials, nor does the agency assume responsibility for their accuracy.*

As the Foundation develops policy and strategy for long-lived digital data, it is essential that the traditional distinction between NSF initiated and disseminated data, and data created, maintained, and shared by its grantees be maintained.