# I. Summary

The Data Archive Task Force (DATF) was convened in the late fall of 2013 as a subcommittee of the Research and Creative Activities Committee (RCA) with the task of considering university policies toward the archiving of research data, including both data accessibility and infrastructure, particularly with regard to upcoming changes in federal policy toward data accessibility for funded research.

Accurate and retrievable data are an essential component of any research project. Such data are valuable to scholars for the duration of their research, and serve the overall mission of the University. Data must be managed to ensure security and compliance with funding agency regulations. Scholars need the right equipment and instrumentation, useful physical facilities, and access to archived information, as well as platforms and tools for collaboration.

A recent survey of faculty and PIs of sponsored research awards shows the University's current research data infrastructure reflects unwarranted fragmentation and a lack of security, both of which could potentially impact the institution's status as a premier research university and expose the institution to undue risk. The ASU research community is often unaware of available capacity, services, and issues surrounding data storage, transport, and sharing. The University should improve coordination, education, communication, and security compliance to potentially realize significant cost savings and enhance opportunities through information, equipment, and instrumentation sharing.

The university is now at a crossroads with tremendous opportunity to grow leadership in data stewardship, lower costs of data resources and management, and move ASU toward a secure and sustainable future in the harvesting and exploration of data. The DATF has taken a forward looking approach to data storage and accessibility and believes the University can be proactive and become an international leader on these issues. We believe general policies and solutions should be generated which broadly apply to **all research and scholarship** generated through the university, regardless of funding source.

The DATF recommends the following actions be taken:

1. The University Senate should establish a one-year Data Archiving Working Group (DAWG) to address and coordinate recommendations in the areas outlined in the remainder of this report. The working group should contain members representing a broad array of disciplines across the university, as well as potential representatives from units such as the University Technical Office (UTO), the ASU Libraries, the Office of Knowledge Enterprise and Development (OKED), the Internal Review Board (IRB), and the Provost's Office. (Section IV.g of the appendix below contains comments and suggestions from ASU Libraries on this report that would be important for the DAWG to consider.)

2. The chairs/directors of all academic and research units should disseminate this document to their faculty and begin an internal discussion over data archiving requirements and

standards for their unit. Many disciplines may already have data archiving standards, while others may need to develop theirs from scratch. Questions which cannot be answered within the unit, as well as feedback on the policies, should be addressed to the Senate working group. Eventually, every unit should generate official data archiving and accessibility policies, consistent with the broader University policies being developed (recommendation 3), to be integrated into unit policy documents.

3. The policies outlined in the remainder of this document are not mean to represent the final University policies, but rather are intended as the suggested starting point to generate discussion. When adequate feedback and discussion have been generated from the university as a whole, an official set of recommended policies should be brought to the University Senate for consideration and formal codification within the ACD and/or other university policy documents, as appropriate.

# Contents


I. Summary ..... 1

II. Introduction ..... 3

II.a. Some General Definitions and Specifications ..... 3

II.b. Data and the University ..... 4

II.c. Recommendations and Plan ..... 5

II.c.i. Recommendations to the University ..... 5

II.c.ii. Plan for a one-year Data Archiving Working Group ..... 5

III. Data Policy ..... 7

III.a. General Principle ..... 7

III.b. Data Archiving ..... 7

III.c. Data Accessibility ..... 8

III.d. Frequently Asked Questions ..... 10

IV. Appendices and Support Information ..... 12

# II. Introduction

## II.a. Some General Definitions and Specifications

- What do we mean by *research* in this document?
    - Unless otherwise specified, the term *research* is being used broadly throughout this document to refer to both **research and scholarship**, including all forms of creative output. The ideas discussed in this document should be considered by all units, regardless of discipline, from the arts and humanities to the sciences and engineering. How these policies apply to specific units can only be determined from within each unit; thus, broad feedback from all units on these policy suggestions is necessary to ensure inclusiveness and responsiveness to the entire university community.

- What do we mean by *data* in this document?
    - Data varies dramatically from field to field and discipline to discipline and can take many forms. The following document is **primarily focused on digitally stored data**, although many of the principles should apply to non-digital data.

- What do we mean by *raw data* in this document?
    - In an optimal world, all data would be archived in its rawest collected form, with intermediate, cleaned-up, and derived data sets also stored with clear protocols and pathways describing how each was derived from the rawest data. For a number of reasons this level of data archiving is unrealistic and unreasonable. The details of the rawest form of data are often extremely specific and unusable until cleaned up into a more summary form that serves as the basis for analysis. The volume of raw data can also be overwhelming: high-energy physics and high-throughput genomics may generate terabytes to petabytes of raw data every day; it is only after these data are combined, averaged, and summarized that they reach a form which could ever be considered useful to the broader community. The term *raw data* in this document should be viewed as the rawest *reasonable* form as defined by the standards of an individual's research community.
    - The goal of these policies is not to define these standards but to encourage that standards be followed. For research communities where standards do not currently exist, units and investigators are encouraged to start a discussion of developing such standards.

## II.b. Data and the University

Central to the University's research mission is accurate and secure data, in many formats, which can be efficiently and safely stored, organized, transported, preserved, and shared. Data must be managed so as to ensure security and compliance with funding agency regulations. Scholars need the right equipment and instrumentation, useful physical facilities, and access to archived information, as well as platforms and tools for collaboration. A high level of confusion and inefficiency is generated in the current under-resourced environment.

Over the course of the DATF discussions on the various services that comprise the University's research support infrastructure, it became clear that the ASU research community is often unaware of available capacity, services, and issues surrounding data storage, transport, and sharing. The DATF recommendations derive from the need for increased coordination, education, communication, and compliance. By addressing these issues, the University can potentially realize significant cost savings and enhance opportunities for policies, governance, information, equipment, and instrumentation sharing.

For example, a recent survey shows ASU's research data is currently stored in scattered locations, including researchers' offices and homes, and departmental servers. This creates an environment where responsibilities for data storage, backup, migration, preservation, privacy protection, compliance, and security can be ambiguous or neglected. For the University's scientific equipment, there is no comprehensive inventory, adequate technical support, or mechanism to track depreciation or fund replacement data equipment. Similarly, content, in the form of databases, educational media, and the University's own scholarship, represents a major University investment, yet a lack of coordination inhibits information sharing and increases costs through duplicate purchases. These are key areas for improvement of research data support at ASU. The current fragmentation and redundancy may represent an opportunity to achieve significant cost savings by leveraging the economies of scale that come from centralizing networking, computing, storage, power, cooling, and staffing. A program that offers efficient, secure, and compliant data stewardship is critical to the University's status as a premier research university and its ability to attract sponsored research funding and premier scholars.

On the upside, ASU has made inroads in the area of network technologies. For example, ASU was one of the early participants in the Internet2 consortium. The primary goals of Internet2 are to create a leading edge network capability for the national research community, enable revolutionary Internet applications, and ensure the rapid transfer of new network services and applications to the broader Internet community. Through its participation in Internet2, ASU has discounted access to a number of important cloud based services. In addition, ASU has recently invested in upgrading the legacy wired and wireless infrastructure throughout the University.

## II.c. Recommendations and Plan

### II.c.i. Recommendations to the University

To address these issues, the University should:

1. Provide, communicate, and enforce policies to ensure data security and compliance (see section III for suggestions).

2. Educate scholars in the existing available research support infrastructure capacities and services.

3. Provide access to services and facilities for the storage, backup, and preservation of research data to allow researchers to meet their ethical and/or legal requirements.

4. Provide access to training and support in research data.

5. Explore ways to reduce fragmentation in the areas of a) data storage, transport, and management; b) purchasing and maintenance of data equipment and contract.

### II.c.ii. Plan for a one-year Data Archiving Working Group

The DATF recommends formation of one-year Working Group to address the following proposals and any others it deems appropriate:

1. Expand the centralization of research data storage, organization, and management, and increase awareness of existing institutional repositories.

2. Perform a comprehensive, continuing analysis of services, needs, opportunities, and costs associated with various research data (and educational data) solutions.

3. Explore greater coordination of the management of research data across colleges and units, and explore sustainable funding models. Possibilities include an All-Arizona consortium (ASU, UA, NAU), a Pac-12 consortium, a Pac-12/Big-10 Consortium, or a more scattered set of education partners.

4. Ensure that research data are stored on centrally-administered computing systems that are managed using up-to-date protocols to ensure data security and compliance with all relevant federal regulations, including those for disaster recovery and continuity.

5. Review, update, communicate and enforce policies for research data security to augment policies for general data security.

6. Promote creation of effective Data Management Plans for all research data generated for sponsored research proposals.

7. Review, update, communicate and enforce policies for research data transport as they apply to both wired and wireless networks.

8. Address funding issues for equipment maintenance, depreciation, replacement, and support.

9. Determine how the policies, once in place, are to be governed. Who will have ultimate oversight of the policies and their implementation? Consider whether the University needs a Data Management Officer (DMO) to oversee, communicate, and enforce these policies. If so, recommend where within the administrative structure this person should reside (obvious possibilities include: Provost's Office, OKED, ASU Libraries, or UTO) and what their responsibilities and charge would be. Should these responsibilities be added onto an existing administrator or is an independent position necessary?

# III. Data Policy

The following are suggested general policies for data archiving and data accessibility. They are not meant to represent final policy, but are intended as the starting point for discussion.

## III.a. General Principle

The guiding principle of this document is that all data generated by research at the university should be archived and, when allowable, made publically accessible. Although related, the issues of archiving and accessibility need to be treated separately since the timing and requirements for archiving may be different than those for making data accessible. A strong university policy on archiving and accessibility of data will aid in securing external research funding as more funding agencies build requirements about data storage and access into the proposal review process.

Although most data generated at the university will be collected by individuals other than the lead investigator, including (but not limited to) students, postdoctoral researchers, and technicians, data archival is still the primary responsibility of the principal investigator, and the following procedures and principles are expected to be followed by faculty in the course of their research.

## III.b. Data Archiving

All faculty are expected to archive their data as a matter of course. There are a number of reasons for this, including (1) data generated by faculty represent Intellectual Property of the university and should be maintained for potential future university use; (2) data used as the basis of publications should be available for reanalysis or examination if questions about a study are raised—**archiving data protects the faculty member against accusations of research misconduct** (including fraud and plagiarism); (3) data archival is fundamental to good research; and (4) many funding agencies now expect data archival to be explicitly part of the research proposal and activity.

Broadly speaking, data can be archived in one of two ways. First, it can be submitted to an existing archive maintained by the community, government, or some journals for data of those types (e.g., Genbank, Dryad, Sloan Digital Sky Survey, or Incorporated Research Institutions for Seismology). Second, data can be archived "locally" on a university archive system. For researches with access to the first option, it is not expected that the same data will be secondarily archived locally as well (when the data submitted to the community archive is processed, one might store raw data locally and the processed data on the community archive, but that will depend on the standards of the individual community).

It is clear that the single least efficient way to manage and archive data is the current practice where the means of archival is left to the discretion of every individual researcher. Thus, while data archiving is the responsibility of each individual investigator, the long-term maintenance of a local archive is the responsibility of the University, likely through the Office of Knowledge Enterprise and Development (OKED). It is expected (and has been acknowledged by OKED) that

the university will either create, join, or buy into an archival system that will meet the needs and requirements of both these procedures and policies and of our faculty; and that all faculty will be given access to this system.

Primary investigators will be responsible for ensuring their data is added to an appropriate archival system; the university will be responsible for providing (in one form or another) an appropriate system that will allow for data to be stored in perpetuity. This avoids problems of setting arbitrary limits on the time horizon for archiving data (10 years? 20 years?) as well as issues about the fate of data which result when researchers leave the university, retire, or pass away.

*Timing of Archival*
In general, data relating to specific publications should be archived no later than the time of publication. Data generated from specific grants should be archived at the conclusion of the grant, even if publications are still pending. All research data should be archived after 3-5 years (specifics to be determined by the DAWG), even when unpublished and generated without an explicit source of funding.

Desired characteristics of archived data:

- Data should be stored in open, easily recognizable and parsable formats. Proprietary file types which may become unreadable in the future should be avoided whenever possible.
- Data may be archived in commonly used compressed formats (e.g., zip which is commonly used on all platforms).
- The data archive should not only include the data but also meta-data about the data set, including a description of the data, a basic explanation of how it was generated, units, and any other information necessary for someone else to comprehend. Reference to an external publication describing the methods would be considered minimally adequate, but copying the relevant information into an associated document within the archive is preferable.

## III.c. Data Accessibility

Beyond storing data in an archive, in general it is preferable for data to be made openly accessible to other researchers and the general public. Again, **these policy suggestions are primarily aimed at digitally stored data.**

- As a public university, the vast majority of our research is funded, both directly and indirectly, through public funds. The public is therefore a stakeholder in our output and should have access to it.
- There are sound scientific reasons for openly accessible data, including (1) verification, (2) repeatability, and (3) reuse. Studies have shown that authors with open data sets get more recognition and citations than those without.

- External funding agencies are moving more and more to requiring open access to data as a requirement for funding.

In general, the following principles about open access apply to all data generated through research at the university, with the following exceptions:

- Classified research
- Proprietary data used in research
- Specific constraints mandated by a funding agency
- Data whose public release would violate clear ethical or legal standards and doctrines

Additional exceptions may be granted by the person or unit ultimately made responsible for governance of the University policies (e.g., the Data Management Officer if the DAWG determines one is necessary; otherwise, perhaps unit chairs and directors).

*Human Subjects Research*

Unless public release of identifiable data is explicitly part of the approved IRB protocols and consent forms, only *de-identified human subject data* should be made publicly accessible. As part of the IRB application and approval process, investigators should work with the IRB to guarantee subjects have informed consent that de-identified data will become part of the public record, including an explanation of the timing of such releases.

Disciplines which include the use of human subjects but which are not traditionally subject to IRB oversight (for example, visual arts such as figure drawing, painting, and photography) need to develop their own policies for whether archived data beyond the finished project (e.g., preliminary sketches or photographs) should be openly accessible or archived with restricted access, keeping in mind the generating principles of these policies as well as ethical, privacy, and safety standards.

*Timing of Accessibility*

The timing of when data becomes open access and accessible is much more complicated than that of general archival. One needs to weigh the benefits of the data to the collecting researchers against those of society. Generally, researchers should be given adequate time to profit from data they have collected, but this should not be abused in such a way as to indefinitely prevent community access.

The following guidelines, in conformation with best practices of an individual's discipline, are expected to be followed, unless superseded by specific requirements of publishers or funding agencies (i.e., one may not use the university guidelines to deny open access to data as required by a publisher).

- It is generally preferable that data be made accessible at the time of publication, but when planned multiple publications will come from the same data set, public access can be restricted until the final paper is accepted for publication (unless required by a journal).

- Unless meeting one of the exceptions noted above, public access to data must not be indefinitely withheld. All allowable data should be made publically accessible within a specified number of years of collection or after funding has ended; the specifics of this timing need to be determined by the DAWG in consultation with individual units.

## III.d. Frequently Asked Questions

- **Question: Does data of type ____ need to be archived?**
  - *Answer: Standards for specific types of data to be archived need to be generated by units in concordance with the standards of their disciplines, should such standards exist. However, this question may best be answered by remembering the two principle reasons why data archiving is important: (1) it protects the investigator against accusations of misconduct, including fraud or plagiarism; and (2) the data has potential for future reusability. If neither of these conditions is likely to be met for a particular type of data, then the requirement to archive is likely low.*

- **Question: How will these policies be enforced?**
  - *Answer: Passively. We anticipate that there will be little active enforcement of these policies within the university, once they and the necessary infrastructure are put into place. Internally, we expect these policies to only come into active enforcement if there is a request (whether internal or external) to see data and it cannot be provided. It is possible that individual units may choose to integrate data archive requirements into their yearly or P&T review process, but at this time we anticipate this to be a unit-by-unit decision. (If the university should choose to pursue more active enforcement, then specific enforcement policies will also need to be generated.) External to the university, many publishers and funding agencies are moving toward stricter requirements on data archiving and release. Individuals publishing with or receiving funds from such institutions will be required to follow such policies as a general means of publishing or receiving future funding.*

- **Question: Should data which can be misused be openly accessible?**
  - *Answer: It depends very much by what is meant by misused. If the misuse is to support a disagreeable assertion, then the open nature of the data is actually a benefit because anyone can go back to the data and demonstrate the misuse. If the misuse crosses legal boundaries, then a solid argument for restricting access likely exists. In general this is something which needs to be determined by individual investigators in consultation with their unit heads and the Data Management Officer.*

- **Question: How do these policies apply to non-digital data?**
  - *Answer: These policies were broadly written for digital data, but many of the same principles apply to non-digital data. The primary differences for non-digital data are: (1) a central archival system is not a realistic option, since forms of physical data*

*would vary tremendously with extremely different archival requirements, thus physical archival must be the purview of individual investigators and/or units; (2) the time horizon for archival of physical data is potentially shorter and must needs be determined by the standards of the discipline; and (3) accessibility takes on an entirely different meaning and, again, must follow the standards of the discipl*ine.

- **Question: What policies should I follow when the university policy is in conflict with an imposed policy from a funding agency or publisher?**
    - *Answer: Generally, one should follow whichever policy is stricter, because by doing so one should meet the needs of both policies. For example, if one agency requires 5 years of archival and another 10 years, you should plan on at least 10 years of archiving. If the university policy is open access, but the funding agent requires that data be kept confidential, then confidentiality must be maintained. Questions about specific conflicts should be addressed to the Data Management Officer (or whoever is designated for this).*

- **Question: This document refers to the standards and policies of my discipline. If my discipline has no such standards and policies, does this mean none of this applies to me?**
    - *Answer: No. While many disciplines have begun to develop standards and policies for data archiving and access, many have not. If you are in one of the latter disciplines, you might view this as an opportunity to shape the future of your discipline by opening a discussion as to what the standards should be. Consider taking some initiative with your colleagues and collaborators and the leadership of your societies and think about how data should best be handled. At the very least, you should plan on developing standards and policies for your local unit, even if those are simply to follow the general University guidelines.*

# IV. Appendices and Support Information

**Contents:**

a. OSTP Memorandum on Access to the Results of Federally Funded Research
b. Report of the NSAC Sub-Committee on Public Access to Research Results
c. Data Archiving in Psychology and Issues of Human Subjects Data
d. Dryad Digital Repository
e. FORCE 11 Draft on Data Citation Principles
f. Research Articles on Data Accessibility
g. ASU Libraries comments on this document
h. Results of an internal ASU survey on Research Data

## IV.a. OSTP Memorandum on Access to the Results of Federally Funded Research

On February 22, 2013, John Holdren, Director of the Office of Science and Technology Policy (OSTP) in the White House, released a memorandum titled "Increased Access to the Results of Federally Funded Scientific Research." The opening paragraph is:

> "The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data."

The full memorandum can be found here: http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

Specific policies, both across and within Federal agencies, are still being developed, but any plan developed by the University will need to be compatible with outcomes deriving from this directive. It is the believe of the DATF that broad policies which encompass all research within the University, whether federally funded or not, would reflect the community spirit and best scientific practices behind this memorandum.

## IV.b. Report of the NSAC Sub-Committee on Public Access to Research Results

The Nuclear Science Advisory Committee (NSAC) released a detailed report on public access to research results in July, 2011, available here: http://science.energy.gov/~/media/np/nsac/pdf/docs/NSAC_PARR_report_final.pdf. The executive summary of the report is:

> One aspect of the America COMPETES Reauthorization act is public access to research results, particularly in the forms of scholarly publications and digital data. In response to this, the DOE Office of Science has charged its advisory committees with identifying and

assessing the current policies, procedures, and practices for disseminating research results; this report is on the research results in fields that are relevant to the Nuclear Physics program.

Finding 1:
The field of nuclear physics publishes in scholarly journals and uses the publication policies of those journals as well as archives and databases to make its research results available to the public. The results available through these means are the peer-reviewed versions of record (VOR). The VOR represent the ultimate product of the government investment in research and are uniformly available to the public. In most cases this access is free, and in others there is a cost associated with access through the journals themselves. Whenever possible, authors make the VOR available at no cost to anyone who requests them. Measurements in the VOR are often used by others to derive additional physics results.

Finding 2:
Pre-final data in the form of preliminary data, theses, conference presentations, and reports are generally publicly available on pre-print servers (e.g. arXiv and CERN Document Server), conference websites, and published proceedings, and, in some cases, in collaboration talk data-bases. Such results are often disseminated in workshops where through collaborative discussion the results are further understood and developed. In some cases the digital data presented in figures are also made available upon request. Requests from the general public for access to pre-final data are not common.

Finding 3:
Requests for digitized detector signals, processed detector signals, and associated computer codes by others not involved in producing them are in general rare, and because of the complexities in using these data, usually not fulfilled. The knowledge and resources required to utilize these data generally make them useless to persons unfamiliar with the experimental apparatus and the conditions under which the data were collected. There have been exceptions where dissemination of such data was useful, and under these situations these data were provided after publication. There are also situations where scientists may join the collaborations processing the data and then participate in the analysis effort.

Finding 4:
Small focused workshops (such as those at the Institute for Nuclear Theory), summer schools, collaboration meetings, and conferences play a crucial role in disseminating and extending research results. A deeper understanding of both experimental and theoretical nuclear science is enhanced by one-on-one interactions in these settings. The dissemination and sharing of pre-final research at these orkshops inspire advances in the field.

## IV.c. Summary of Data Archiving in Psychology and Issues of Human Subjects Data

- The APA provides links to Shared Data Sets and Data Repositories available for psychological science research: http://www.apa.org/research/responsible/data-links.aspx
- ICPSR (Interuniversity Consortium for Political and Social Research) is an example of a social science repository; ICPSR advances and expands social and behavioral research, acting as a global leader in data stewardship and providing rich data resources and responsive educational opportunities for present and future generations. http://www.icpsr.umich.edu/icpsrweb/landing.jsp
- APS published some articles on data sharing in 2002 in their professional magazine (*The Observer*). The lead article notes that the field of psychology has been slow to publish data and that there is no reason for this as there are no obstacles to data sharing that cannot be overcome. http://www.psychologicalscience.org/observer/0102/databases.html
- APS has announced new initiatives to improve publication standards and practices for their journals. http://www.psychologicalscience.org/index.php/publications/journals/psychological_science/ps-submissions
    - Among these is "Promoting Open Practices" that will associate "badges" with articles "to acknowledge their authors have made the data, materials, or preregistered design and analysis plans publicly available. Badges will be awarded to qualified articles, which must include links to the open data, materials, or preregistered design and analysis plans."
    - An associated initiative is the use of "New Statistics" to make results fully available regardless the outcome (e.g., counteracting the fact that null results tend not to be published) and to ensure the integrity (completeness, coherence, trustworthiness) of publically available research. The New Statistics calls for an end to NHST (null hypothesis significance testing) and advocates use of estimates based on effect sizes, confidence intervals and meta-analyses.
    - This "New Statistics" movement extends beyond psychology to NIH, epidemiology, and other scientific disciplines in which it is believed that much of the publically available research is wrong. http://pss.sagepub.com/content/early/2013/11/07/0956797613504966.full
- There have been several workshops and reports on the topic coming out of the National Research Council (NRC).
    - In 2012 the NRC Board on Research Data and Information released a report that summarizes an International Workshop held in that year. The report is titled "For Attribution – Developing Data Attribution and Citation Practices and Standards." http://www.nap.edu/catalog.php?record_id=13564
    - The NRC Division of Behavioral and Social Sciences and Education held a public comment meeting last year ( May 16-17, 2013) concerning access to Federally Supported R&D Data http://sites.nationalacademies.org/DBASSE/CurrentProjects/DBASSE_082378

        - The agenda for this meeting is at: http://sites.nationalacademies.org/xpedio/groups/dbassesite/documents/webpage/dbasse_083128.pdf
    - Among many issues discussed at that meeting were some commentaries on data from human research. A representative from AERA stated that "OSTP, federal agencies, and the Office for Human Research Protections should develop a statement to foster responsible sharing of identifiable as well as linked data as long as scientists use such data under restricted conditions, are legally bound to honor consent agreements, and face stringent penalties for disclosure. The NRC, federal agencies, data repositories, or scientific societies could assist in this task."
        - Bob Hauser (DBASSE Director) commented that the Office for Human Research Protection (US Dept. of Health and Human Services) is undergoing changes to the 45CFR46 "Common Rule." The federal government is contemplating various ways to enhance the regulations overseeing research on human subjects.
        - Interestingly, some of the changes under consideration concern data security and gaining consent by human subjects to use even de-identified data. http://www.hhs.gov/ohrp/humansubjects/anprmchangetable.html
        - There is a more recent NRC report on proposed revisions to the Common Rule which includes discussion of a meeting session dedicated to Data Use and Sharing and Technological Advancements. http://www.nap.edu/catalog.php?record_id=18383

**Conclusion.** Although the field of psychology has not led the way on data archiving, there is now significant movement in that direction. There are a number of data repositories for the field and changing journal practices that encourage making data publically available. Further changes in the Common Rule governing human subjects protection seem to be anticipating issues associated with making human subjects data publically available.

## IV.d. Dryad Digital Repository

The Dryad Digital Repository (http://datadryad.org) is one example of an existing general data archiving service, governed by a nonprofit membership organization. It is being presented here simply as an example of one approach to solving many of the data archiving issues discussed. The community of Dryad consists of (http://datadryad.org/pages/organization):

> Data are the foundation of the scientific enterprise. By preserving and making available the data underlying the scientific and medical literature, Dryad provides benefits to individual researchers, educators and students and to a diversity of stakeholder organizations.
>
> **Researchers:** Dryad benefits researchers by providing free access to data they can use for a host of different purposes: to study and validate published results, for

methodology development, for meta-analyses, and to address novel questions using historic observations. Submitting to Dryad helps researchers get more credit for their work by increasing the usability and citability of their data.

**Educators and students:** Dryad provides educators and students with the opportunity to teach and learn a broad array of analysis techniques, and general data literacy skills, using both classic and recently published research data.

**Scientific societies:** Archiving data in Dryad strengthens the legacy of a scientific society by permanently preserving the data and increasing the research impact of its members and community. The availability of data creates new opportunities for research and education and promotes public trust in science.

**Journals and publishers:** Dryad frees journals and publishers from the responsibility and costs of publishing and maintaining supplemental data in perpetuity. By encouraging a broader range of reuse, journals and publishers increase the impact, citations and the prestige of their publications.

**Research institutes and libraries:** Dryad provides institutions with a new method their researchers can use to showcase their work. It provides infrastructure for the preservation and dissemination of research data collected at the institution, as well as data from other institutions of value to its patrons.

**Research funding organizations:** Dryad provides a cost-effective mechanism for free, long-term access to data which, in turn, enables new research. Data availability improves the rigor of the scientific record and public trust in the scientific enterprise.

## IV.f. FORCE11 Draft Declaration on Data Citation Principles

FORCE11 (http://www.force11.org/) is a nonprofit organization consisting of "scholars, librarians, archivists, publishers and research funders that has arisen organically to help facilitate the change toward improved knowledge creation and sharing." In late 2013, they published a draft Declaration on Data Citation Principles (http://www.force11.org/datacitation) under the following Preamble:

> "Sound, reproducible scholarship rests upon a foundation of robust, accessible data. For this to be so in practice as well as theory, data must be accorded due importance in the practice of scholarship and in the enduring scholarly record. In other words, data should be considered legitimate, citable products of research. Data citation, like the citation of other evidence and sources, is good research practice."

Many of the principles in this declaration directly relate to issues of archiving and accessibility.

## IV.f. Research Articles on Data Accessibility

1. Piwowar HA (2011) Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data. *PLoS ONE* 6(7): e18657. doi:10.1371/journal.pone.0018657

**ABSTRACT:** Many initiatives encourage investigators to share their raw datasets in hopes of increasing research efficiency and quality. Despite these investments of time and money, we do not have a firm grasp of who openly shares raw research data, who doesn't, and which initiatives are correlated with high rates of data sharing. In this analysis I use bibliometric methods to identify patterns in the frequency with which investigators openly archive their raw gene expression microarray datasets after study publication.

Automated methods identified 11,603 articles published between 2000 and 2009 that describe the creation of gene expression microarray data. Associated datasets in best-practice repositories were found for 25% of these articles, increasing from less than 5% in 2001 to 30%–35% in 2007–2009. Accounting for sensitivity of the automated methods, approximately 45% of recent gene expression studies made their data publicly available.

First-order factor analysis on 124 diverse bibliometric attributes of the data creation articles revealed 15 factors describing authorship, funding, institution, publication, and domain environments. In multivariate regression, authors were most likely to share data if they had prior experience sharing or reusing data, if their study was published in an open access journal or a journal with a relatively strong data sharing policy, or if the study was funded by a large number of NIH grants. Authors of studies on cancer and human subjects were least likely to make their datasets available.

These results suggest research data sharing levels are still low and increasing only slowly, and data is least available in areas where it could make the biggest impact. Let's learn from those with high rates of sharing to embrace the full potential of our research output.

2. Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLoS ONE* 2(3): e308. doi:10.1371/journal.pone.0000308

**ABSTRACT: Background:** Sharing research data provides benefit to the general scientific community, but the benefit is less obvious for the investigator who makes his or her data available. **Principal Findings:** We examined the citation history of 85 cancer microarray clinical trial publications with respect to the availability of their data. The 48% of trials with publicly available microarray data received 85% of the aggregate citations. Publicly available data was significantly ($p = 0.006$) associated with a 69% increase in citations, independently of journal impact factor, date of publication, and author country of origin using linear regression. **Significance:** This correlation between publicly available data and increased literature impact may further motivate investigators to share their detailed research data.

3. Piwowar HA, Vision TJ. (2013) Data reuse and the open data citation advantage. *PeerJ* 1:e175 http://dx.doi.org/10.7717/peerj.175

**ABSTRACT: Background.** Attribution to the original contributor upon reuse of published data is important both as a reward for data creators and to document the provenance of research findings. Previous studies have found that papers with publicly available datasets receive a higher number of citations than similar studies without available data. However, few previous analyses have had the statistical power to control for the many variables known to predict citation rate, which has led to uncertain estimates of the "citation benefit". Furthermore, little is known about patterns in data reuse over time and across datasets. **Method and Results.** Here, we look at citation rates while controlling for many known citation predictors and investigate the variability of data reuse. In a multivariate regression on 10,555 studies that created gene expression microarray data, we found that studies that made data available in a public repository received 9% (95% confidence interval: 5% to 13%) more citations than similar studies for which the data was not made available. Date of publication, journal impact factor, open access status, number of authors, first and last author publication history, corresponding author country, institution citation history, and study topic were included as covariates. The citation benefit varied with date of dataset deposition: a citation benefit was most clear for papers published in 2004 and 2005, at about 30%. Authors published most papers using their own datasets within two years of their first publication on the dataset, whereas data reuse papers published by third-party investigators continued to accumulate for at least six years. To study patterns of data reuse directly, we compiled 9,724 instances of third party data reuse via mention of GEO or ArrayExpress accession numbers in the full text of papers. The level of third-party data use was high: for 100 datasets deposited in year 0, we estimated that 40 papers in PubMed reused a dataset by year 2, 100 by year 4, and more than 150 data reuse papers had been published by year 5. Data reuse was distributed across a broad base of datasets: a very conservative estimate found that 20% of the datasets deposited between 2003 and 2007 had been reused at least once by third parties. **Conclusion.** After accounting for other factors affecting citation rate, we find a robust citation benefit from open data, although a smaller one than previously reported. We conclude there is a direct effect of third-party data reuse that persists for years beyond the time when researchers have published most of the papers reusing their own data. Other factors that may also contribute to the citation benefit are considered. We further conclude that, at least for gene expression microarray data, a substantial fraction of archived datasets are reused, and that the intensity of dataset reuse has been steadily increasing since 2003.

## IV.g. ASU Libraries comments on Data Archive Document

*Following is the unedited response of ASU Libraries to the penultimate draft of this document. It is included as an appendix because it contains a number of suggestions and ideas which should be actively considered by the proposed data archive working group.*

---

ASU Libraries comments on "Data Archive TF 2014-02-09" document
February 14, 2014

The ASU Libraries applauds the University Senate Data Archive Task Force for their timely and thoughtful treatment of this important topic. We fully support the adoption of the principles and best practices outlined in this draft policy.

In addition, we ask the Task Force members to consider the following additional comments:

1. The efficiency and long-term effectiveness of Research Data Management at ASU will be enhanced if a Data Management Officer (DMO) position is created. The DMO would ideally be part of OKED but work closely with Data Management staff at the ASU Libraries (Page 6, bullet point 9).
2. Research data management should take into consideration the entire research data lifecycle – that is, from planning, through grant writing, active research, project completion and eventual archiving.
3. Archiving data in perpetuity may be fiscally impossible. Archived data should be managed similarly to other information resources. Resources that are not in demand (analogous to books that have not been checked out in years) are placed in a "dark archive" which is less expensive to run.
4. Research data management is quite different from records management, and different standards and practices are followed. We do need policies and procedures for both but these should be treated separately (Page 5, bullet point 4 – remove reference to records management).
5. We suggest the creation of an ASU Research Data Registry or catalog. This simple listing (managed by OKED or the Libraries) keeps minimal, citation-type descriptive information (project name, PIs, dates, abstract, data files, and LOCATION WHERE THE DATA ARE ARCHIVED). Allowing for multiple data archiving options is important, and provides key flexibility for faculty and PIs. However, we risk losing track of the location of the intellectual capital produced at ASU. A Registry (with a pointer to where the data are archived) allows faculty to archive their data either locally (their own computer), in the ASU repository, nationally or internationally, and maintain a record of the location of that data.
6. Graduate students (Masters and Doctoral) are expected to follow the same (or similar) best practices for archiving (and/or publishing) their research data. Faculty advisors are jointly responsible with the students for ensuring that research data are managed properly. (Page 1, paragraph 4).

---

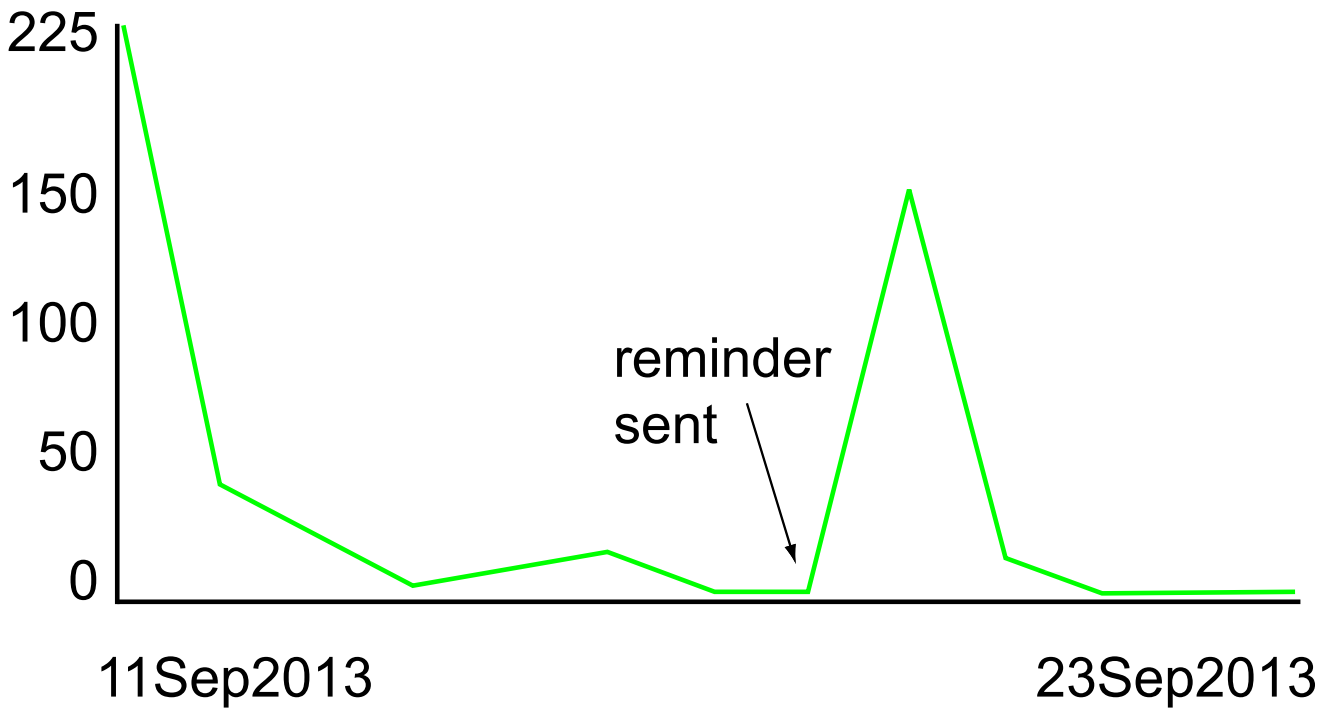
## IV.h. Results of an internal ASU Survey on Research Data

In mid-September 2013, A2C2, OKED, and the Provost's office conducted a university-wide, 17 question survey on how researchers currently store, serve, and secure research data. This information was gathered to provide input on addressing research data capabilities across ASU. The questions asked and the responses to those questions are presented on the following pages.

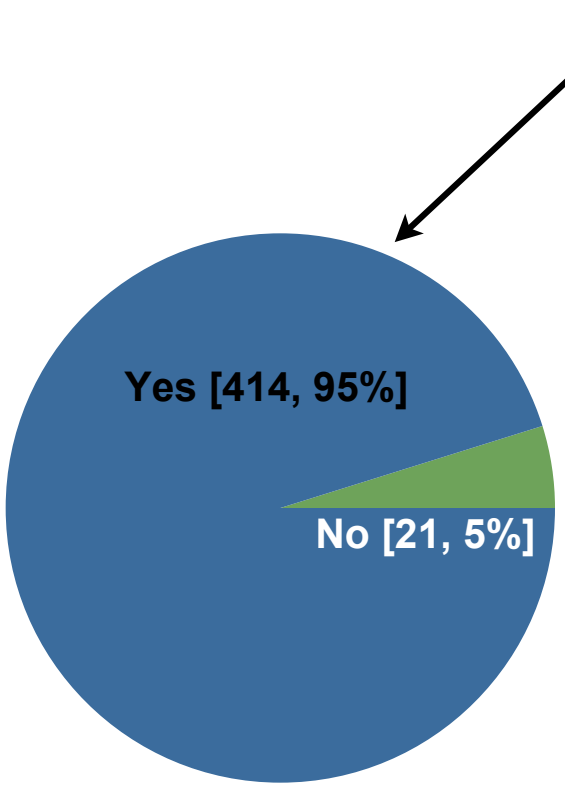
# Results from the research data survey

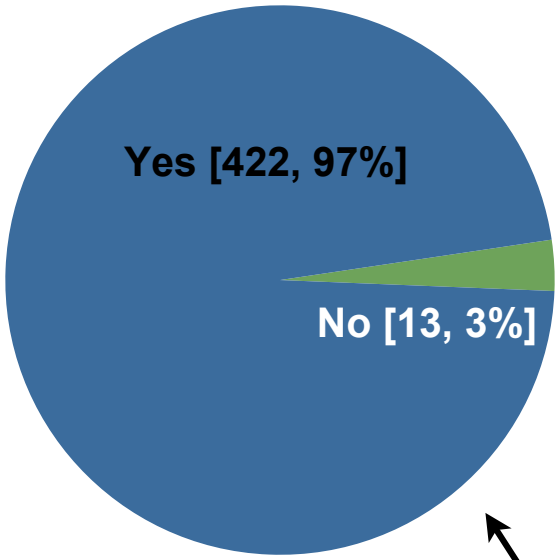
2226 unique email messages sent to tenure track/tenured faculty and PIs of sponsored research awards.

438 responses were received, or19.7% of the messages sent.



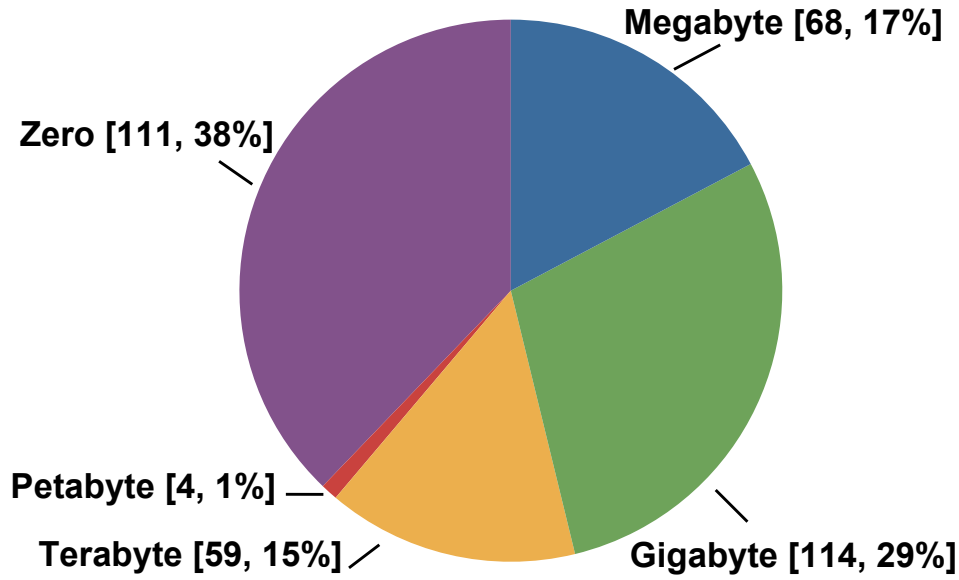
Q1: Do you conduct research that generates some type of digital data (text, images, videos, audio, databases, spreadsheets, instrument files, photographs, physical samples/specimens, etc)?

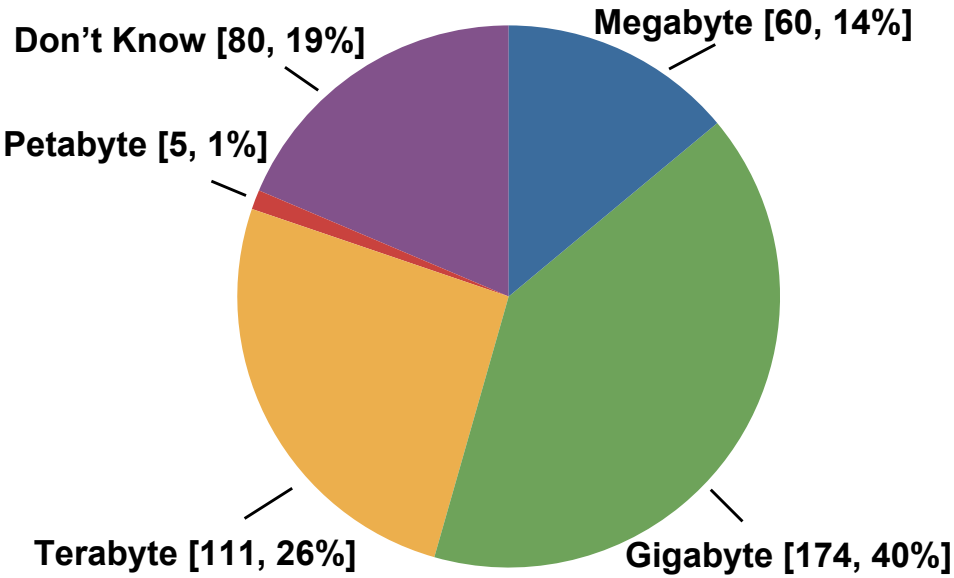



Q2: Are you planning to conduct research that generates some type of digital data in the near future?

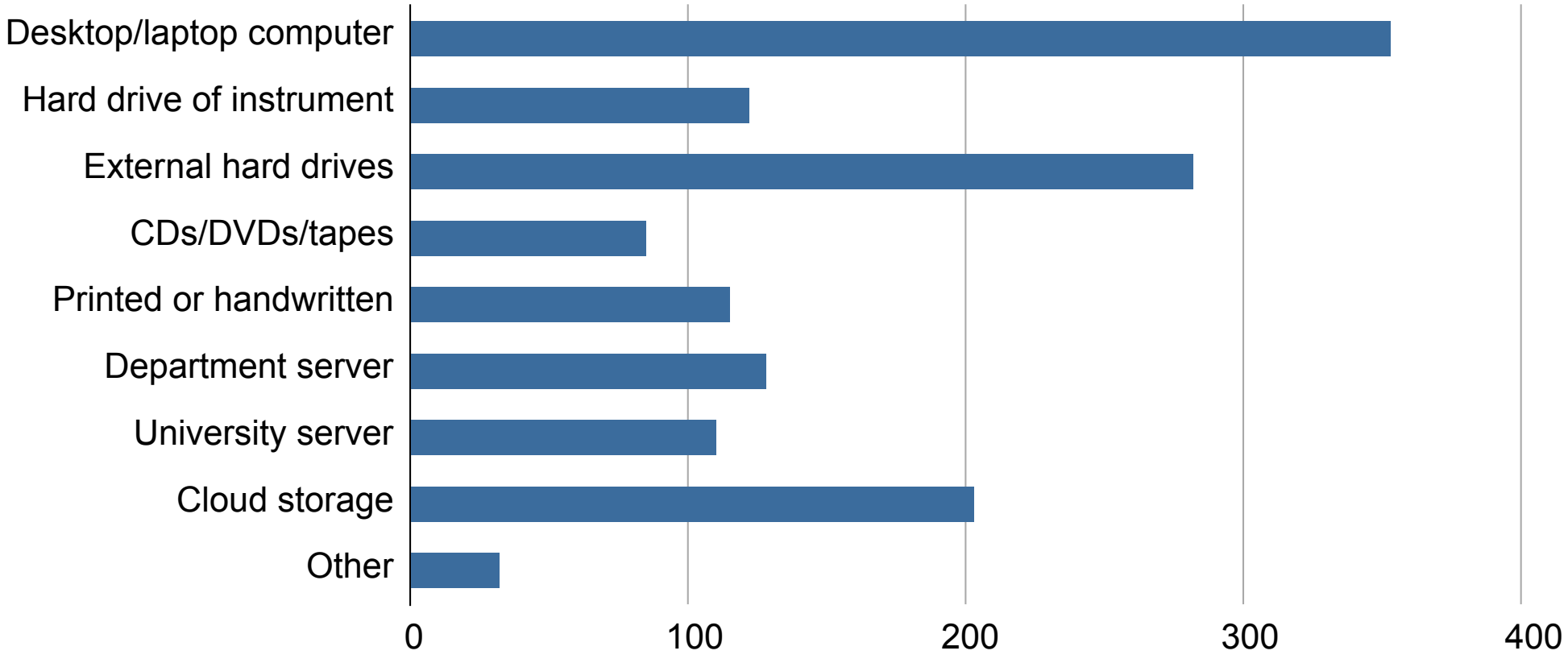
# Q3: Approximately how much digital research data are you currently storing locally at ASU?



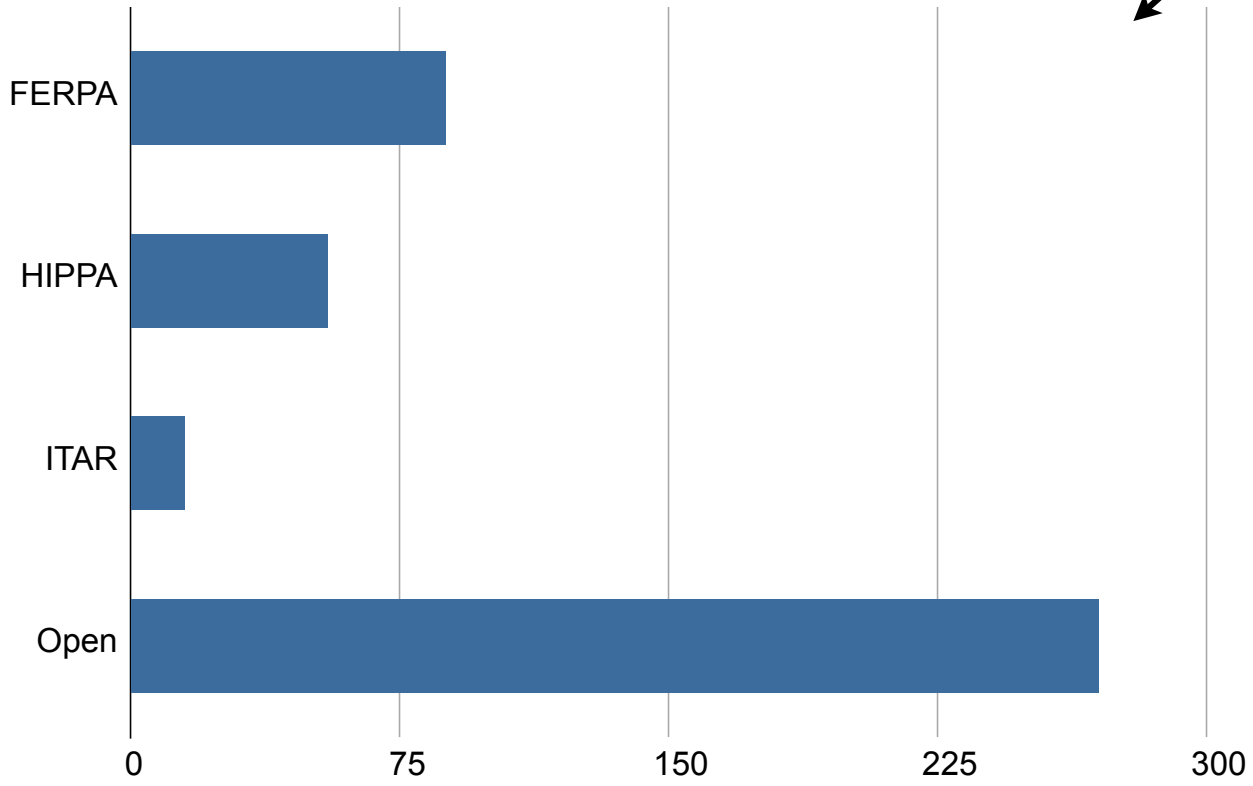



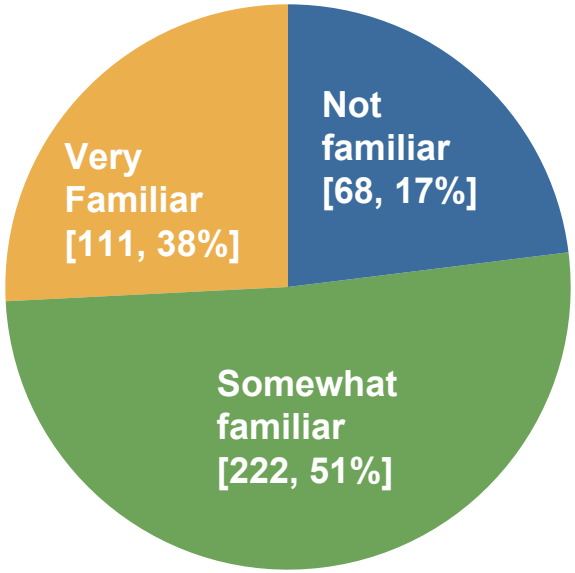

# Q4: Approximately how much digital research data are you currently storing external to ASU?

# Q5: What is your current method of storing research data locally at ASU? (Choose all that apply)



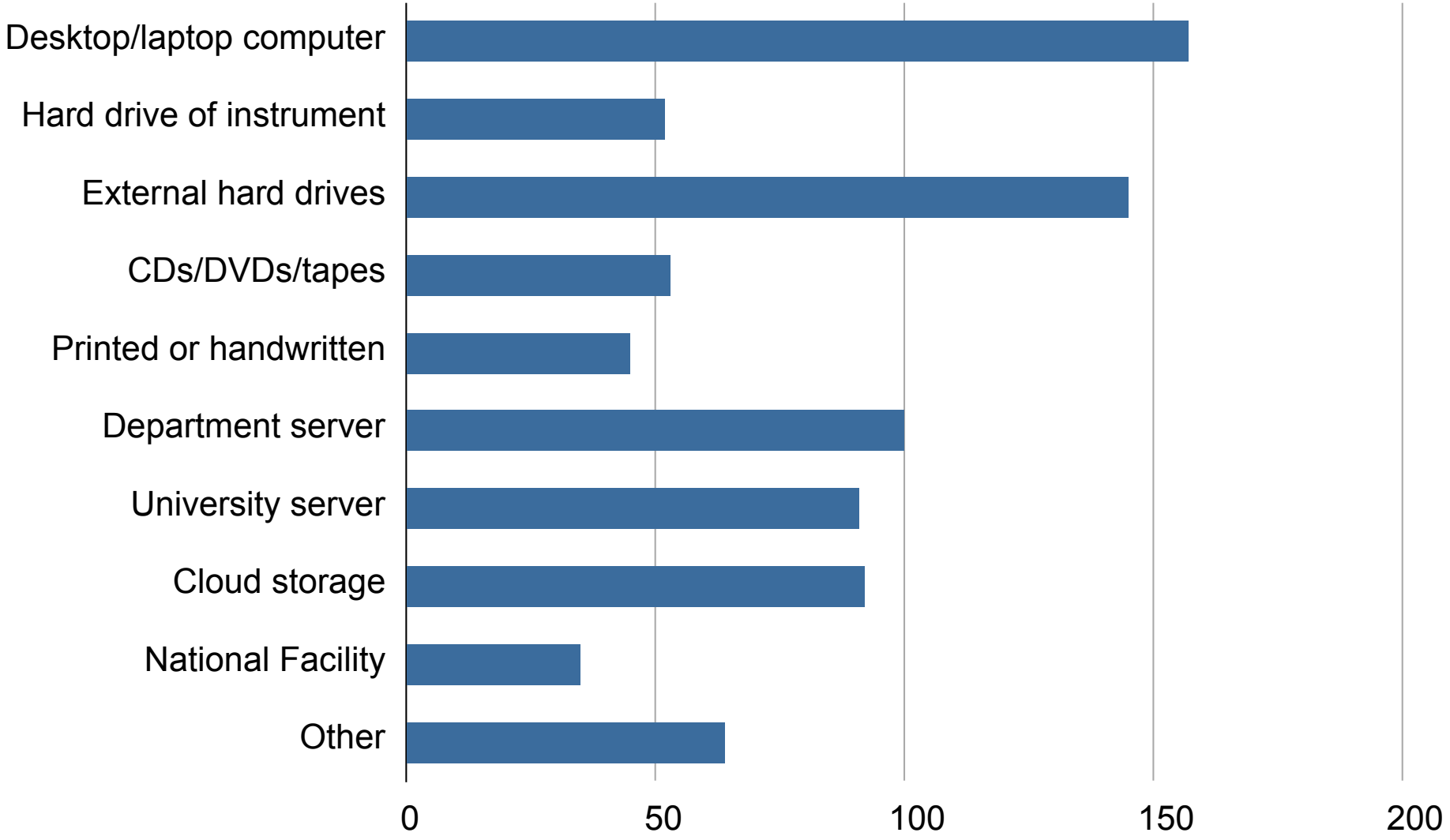
Q6: Which of the following security models applies to your research data?



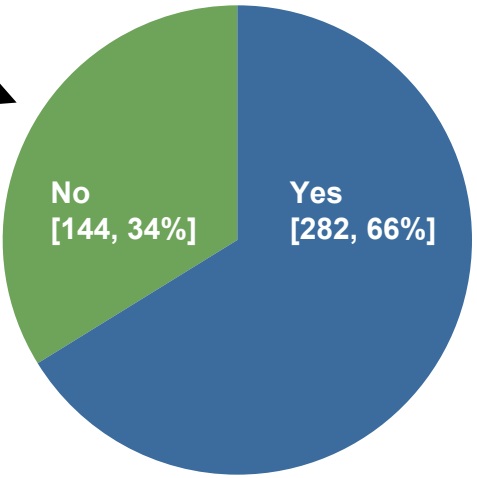



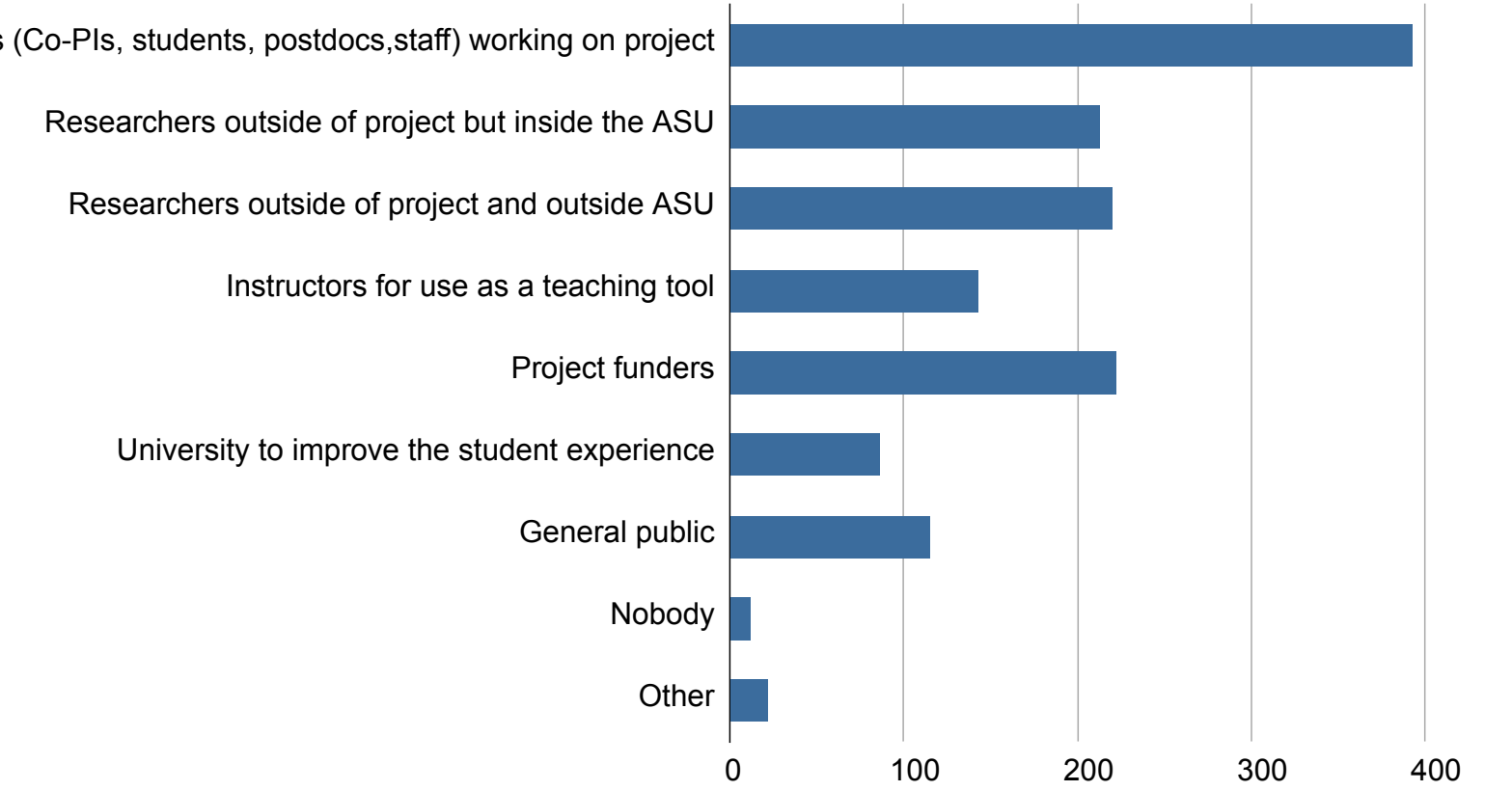

Q7: How familiar are you with the requirement for research data management plans as components of grant submissions to funding agencies (e.g., NSF, NIH, NEH)?

# Q7: For NSF, NIH or NEH awards, how do you, or plan to, preserve and provide access to the research data after award ends?



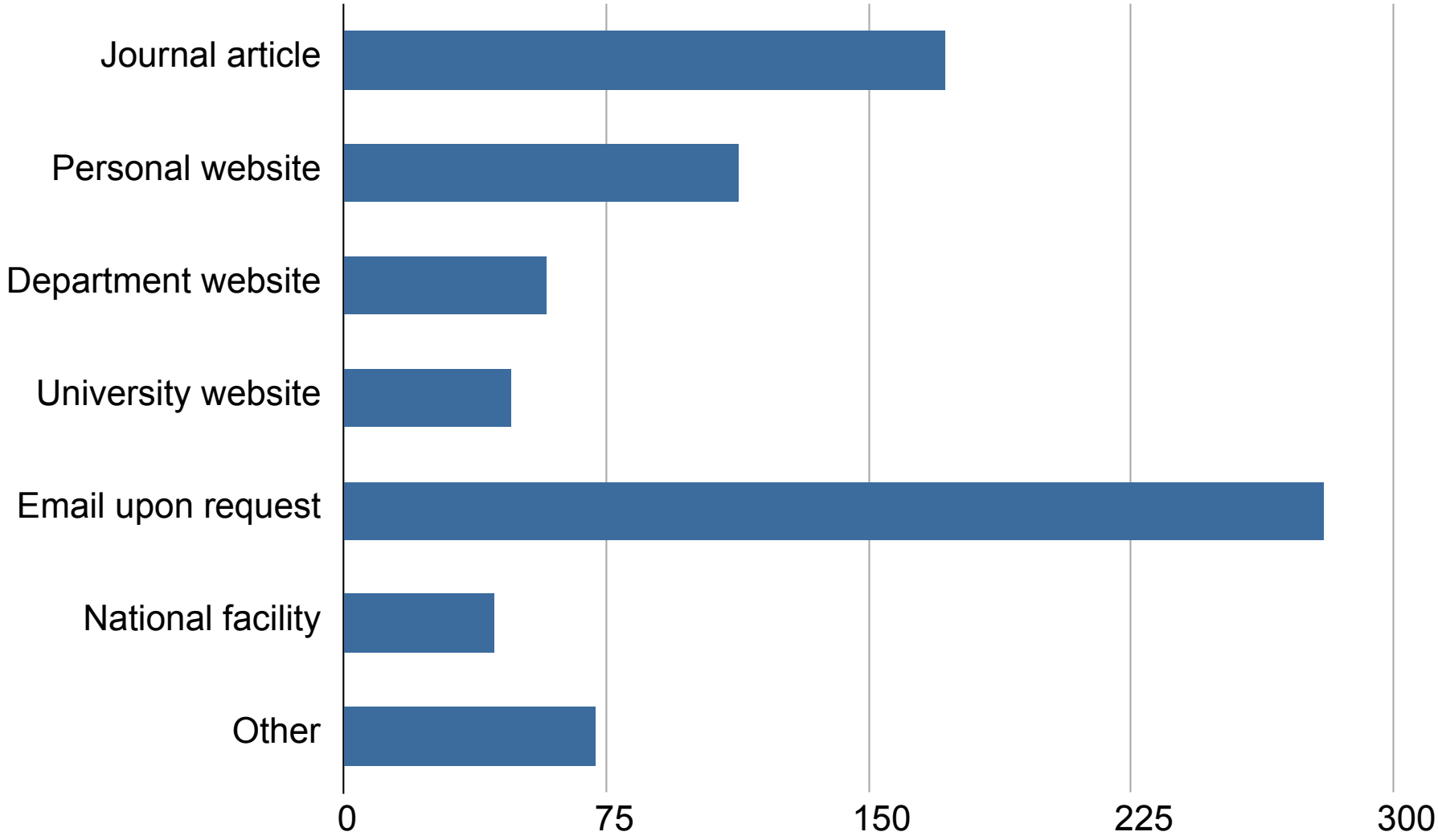
# Q9: Do you currently share your research data with people outside of the University?



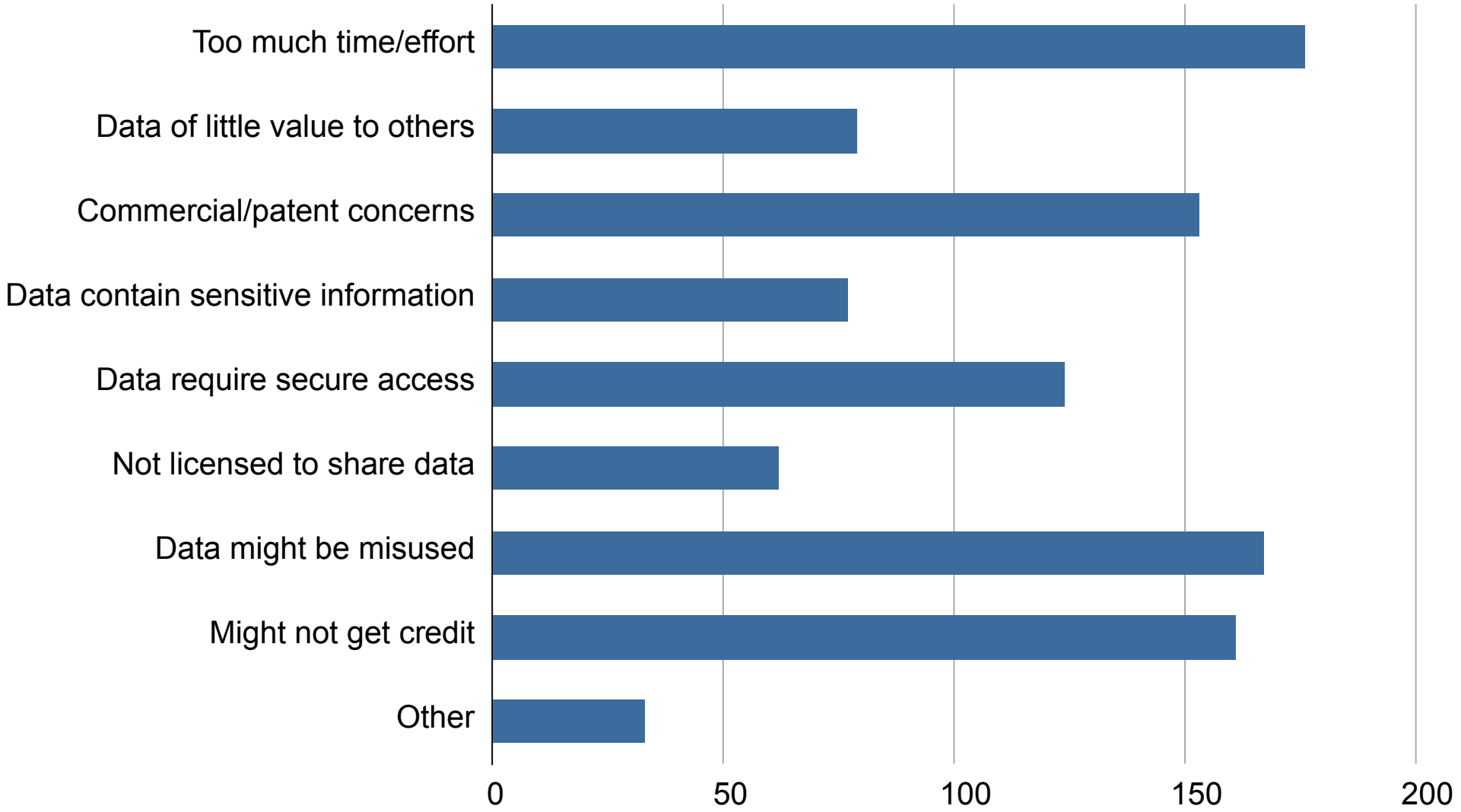




# Q10: With whom are you willing to share your research data? (Choose all that apply)

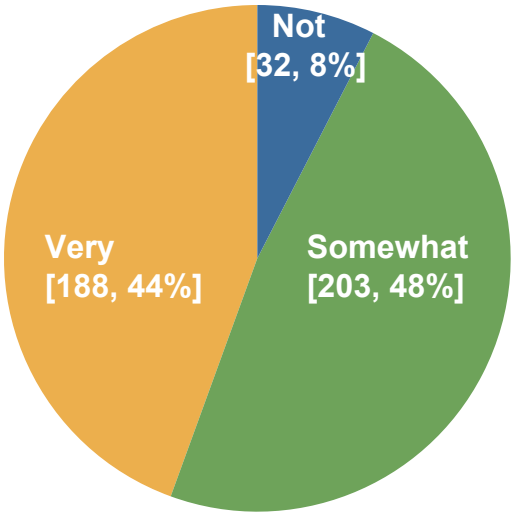
# Q11: How do you share/serve your research data with others? (Choose all that apply)

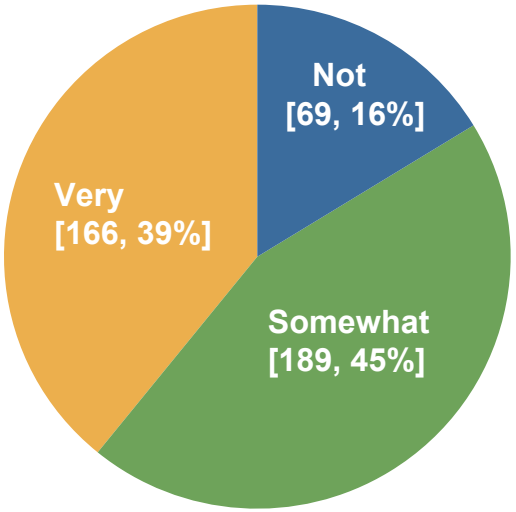


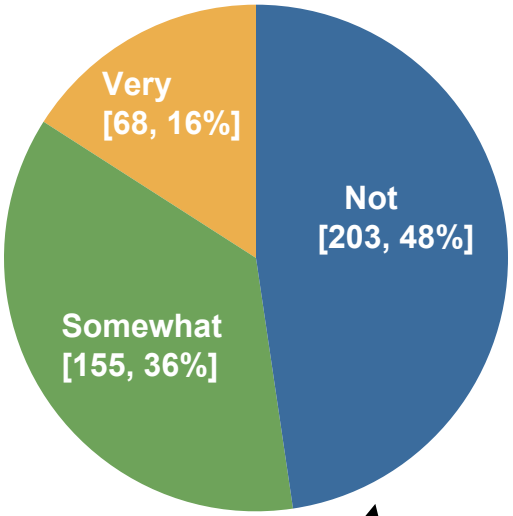
# Q12: What reasons might inhibit you from sharing research data with people outside of your research group? (Choose all that apply)



Q13: How interested are you in storing your research data in a local ASU facility?



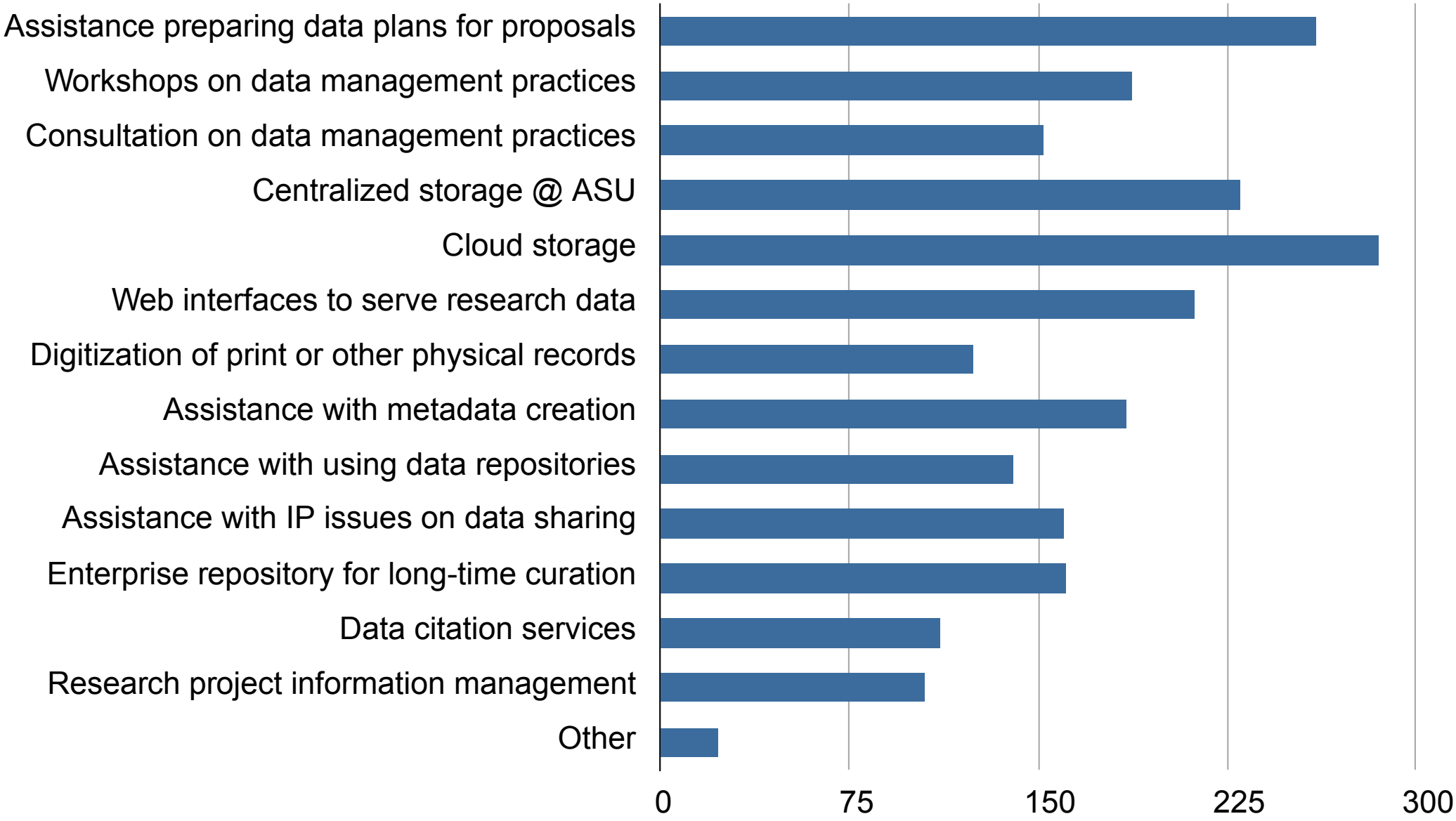




Q14: How interested are you storing your research data in a cloud repository?

Q15: How familiar are you with documenting and/or creating metadata for your data (i.e. so the contents of datasets can be understood by others)?

# Q16: Which research information management services would you use if they were offered by ASU? (Choose all that apply)



# Q17: Academic Unit/center?

197, or about 50%,
responded to this question.