Data Security Standards: Integrity and Availability

Introduction

The Data Preservation Alliance for the Social Sciences (Data-PASS) was formed to take shared responsibility for the long-term accessibility to social science datasets that are of value to current and future researchers and policy-makers. To maintain accessibility and support dissemination, the quality and integrity of the information within and about a data collection must be controlled throughout the various stages in its life-cycle. This document outlines our standards for the security of materials acquired for the Data-PASS project.[1] Based on current security procedures already in place at each organization, these standards protect against the destruction and loss of the data, whether through natural disasters, fire, vandalism and/or error.

Data security is usually understood to involve availability (e.g. through redundancy and management of the computing environment), integrity (e.g., through backups and verification methods), and controlling access (by authenticating users, and authorizing actions on the data). In the archival context, we include data migration within "security", since we use migration to ensure the availability or the intellectual content of the data we maintain, as well as to maintain its integrity.

Data security is closely related both to confidentiality (which includes de-identification and relies upon access control), and to digital rights management (which provides a framework for authorization). This document is one of three that together, address all these issues. This current document discusses the practices that we have identified to maintain data integrity and availability. The associated documents address standards regarding access control; rights management and authorization; sensitive data, de-identification, and confidentiality.

<u>**Inter-Archive Redundancy: Emerging Best Practices**</u>

Each of the Partners maintain unique holdings under the auspices of Data-PASS. In the past, some duplication of archival collections occurred (e.g., both ICPSR and Roper have copies of polling data from ABC, CBS, and the Voter News Service). Unfortunately, this duplication sometimes lead to confusion for researchers, since data versions, identification schemes, and organization were not consistent. Current best practice is moving towards a more fully documented approach to data duplication, which includes maintaining consistent unique identifiers for each resource, and explicit metadata describing the resources, provenance, version, and associated rights. Best practice also is moving towards more systematic and explicit duplication policies, including mirroring of entire collections (rather than ad-hoc selections of individual items), and a process of

---

[1] These standards are intended to reflect best practices among academic data archives. Government agencies, may be governed by regulation. In particular, NARA, a current Data-PASS partner, is governed by: *NARA Code of Federal Regulations*, 36 CFR Part 1234, Electronic Records Management, Subpart C --Standards for the Creation, Use, Preservation and Disposition of Electronic Records

regularly updating a mirror in order to preserve both the original and newer versions of a selected collection.

System Security

In addition to physical security, we need to ensure that the systems we use are also protected. This requires attention to prevention, detection, and response. We use a variety of techniques (such as those identified by CERT -- http://www.cert.org/) to protect the systems that serve our data, including: encryption for logins, secure private networks for administration, automated vulnerability scans, and automated intrusion detection and monitoring. Logs of all system activity will be kept.

Systems Backup

To maintain accessibility of the data and protect against accidental loss, each archive institutes a regular system backup procedure. A combination of full and incremental backups are performed regularly, and stored locally, either on-line (e.g. on a disk mirror) or on a readily accessible medium, such as Digital Linear Tape (DLT).

These backups are kept in a physically secured area, following best practice for data centers such as those identified by Sun and TIA (Telecommunications Industry Associations), and following relevant physical security standards, such as NFPA (National Fire Protection Association), FIPS, NIST, ASHRAE. This backup is stored on-site for quick restoration of the data, as needed.

In addition to the copies of the media kept on the premises, each Partner will keep copies of backup media off-site. The off-site location will be a storage and retrieval facility designed for just this kind of long-term preservation, and provide climate control, physical security, and access control.

All data files, identifiers, and associated metadata will be included in the both the on-site and off-site backups. Backups are monitored for completion status and data verification. All backup media are then indexed and recorded in a database that is maintained at the Partner level. This database, along with cryptographic checksums and universal numeric fingerprints (UNF's), will be preserved for future data validation. .

As a further protection, a shared set of backup systems will be implemented so that all the content acquired will be stored at two sites, ICPSR and Harvard for non-Federal materials and at NARA and ICPSR for Federal records. Periodically, the Partners will update the media transport and replacement schedule in order to maintain the continued quality of the backups and the efficiency of data restoration.

The physical media on which digital data are stored can be damaged, degrade or become obsolete over time. Digital data and the storage media must be checked for readability

and accessibility at regular intervals and, if necessary, copied to new storage media to prevent their loss. Periodic data restoration tests, using random sampling, will be conducted to validate the contents of the media against the database of signatures and UNF's.

Data Migration

Access to data through our storage media is dependent on both hardware and software. To ensure continued access to our data collections through time and changing technology, periodic reviews will be conducted to determine when migration is needed. The Partners will migrate data from any storage media and associated hardware that are becoming obsolete to new media and hardware that are or are becoming standard or currently considered best practice. In making our decisions, we will want to consider criteria such as longevity and viability of the new media. Susceptibility to physical damage should also be considered. In the process of making these transformations, we will also want to protect against the modification of the intellectual content of the original document.

The danger of unwanted modification to the intellectual content is an even greater concern when the migration of data involves new formats. Just as data can become obsolete due to the media on which it is stored, it can also become obsolete if the format in which it is stored becomes unusable. Staff at each of our partner organizations are well-versed in the changes to data and meta-data software and applications. Although many updates allow for the retrieval of data stored in earlier versions, there are occasions when that is no longer possible. Data stored on versions on the verge of obsolescence, will be migrated to the latest version available or current standard formats.

The Partners will conduct an annual review of the media and hardware, as well as the format and software that are used in our data backup and storage, and will migrate data as necessary. As lead partner, IPCSR will record the data status review and migration schedule, so that important maintenance dates for our data will be monitored and proper schedules will be maintained.

These guidelines provide a level of current protection as agreed to in our Cooperative Agreement with the Library of Congress for the content collected under the project funded as part of NDIIPP. The Data-PASS partners recognize that their collaboration allows them to plan collaboratively for future enhanced assurance of data integrity and availability, through adoption of new systems for redundant storage of files, monitoring of data storage, refreshing of data files to new storage media, assessment of format obsolescence, migration of files to new formats, and other means. The rapid rate of technological development will necessitate regular reviews of these procedures.