Idaho State University's Geospatial Coordinating Committee
Best Practice Guidelines for Data Retention and Deletion

Management of digital geospatial data has special requirements as indicated by the *Idaho State Board of Education Records Management Guide* (section 4, special requirements). This document sets forth a policy for the retention and deletion of digital geospatial data for research projects of the ISU GIS Training and Research Center. The policy presented here also serves as a recommendation to the broader ISU geospatial community on best practices for the management of digital geospatial data associated with sponsored research. These best practices conform to the Idaho State Board of Education (SBOE) guidelines as well as American National Standards Institute (ANSI) and International Organization for Standardization (ISO) standards.

Terminology used in this document

- *Active geospatial dataset*: A geographic information system (GIS) or remote sensing (RS) file or collection of files used in normal business activities. These files are typically research or framework data sets stored in a standard geospatial data format. Global positioning system rover files (field data collected by researchers using GPS receivers) in raw or corrected formats are not considered active geospatial data sets until they are converted into a standard geospatial data format. Global positioning system base station files for the current calendar year are considered active geospatial data sets and follow the retention and deletion schedule described below.
- *Archiving*: the process of creating a duplicate copy of an electronic record for future reference. The permanent duplicate copy is stored off-site on a stable external media.
- *Georeference requirements*: Each geospatial dataset should be projected and defined following the appropriate regional standard georeference system. For example, geospatial data for the state of Idaho, in accordance with the Idaho Geospatial Office standard for georeferencing, shall be properly projected and defined in Idaho Transverse Mercator. Data quality (i.e., horizontal positional accuracy) is not addressed by this policy.
- *Geospatial data set*: A GIS or RS file or collection of files.
- *Inactive geo-spatial dataset*: A GIS or RS file or collection of files that is no longer used in normal business activities. These files typically represent a collection of files generated during research that are intermediary to a completed product. These datasets should be stored in a *standard geospatial data format* (see below). In addition, global positioning system base station files for the previous calendar year are considered inactive geospatial data sets and follow the retention and deletion schedule described below.
- *Original data source*: A GIS or RS file or set of files stored in a data format that may or may not comply with the standard geospatial data format guideline. These files represent the status of the data as received. These data typically require various degrees of preliminary processing before they are ready for use in a GIS. Examples include non-georegistered and uncorrected satellite imagery.
- *Standard geospatial data format*: Vector data should be stored as shape files (*.SHP and all associated files that constitute a shapefile dataset). Raster data should be stored in an uncompressed format to include, but not limited, to ERDAS Imagine (*.IMG), ESRI floating point (*.FLT), or tagged image file format (*.TIF),

Data Retention Determination for Archiving

Geospatial data is an asset that should be treated and managed similar to cash or capital equipment. Apart from financial value, it is critical that geospatial data be managed properly to avoid loss due to versioning, corruption, or deletion. Additionally, geospatial data must be managed as a historical dataset for temporal change detection analyses and other purposes. Each Principal Investigator, with guidance from the GIS Director, is responsible for managing geospatial data over the life cycle of that specific data set (generation or acquisition through inactive geospatial data archiving).

Not all geospatial data produced at Idaho State University will be permanently archived. Due to the fact that geospatial data sets are large (many datasets exceed 1GB) it is important to recognize that permanently archiving all geospatial data (i.e., all versions, iterations, and preliminary processing data sets) is not feasible. Instead, geospatial data will be permanently archived to preserve:

- The original digital data source.
- Original digital data that has been processed into a standard geospatial data format.
- Final geospatial datasets stored in a standard geospatial data format, with proper georeferencing, and metadata documentation. These data represent the completion of a model, deliverable, or data layer. All metadata shall follow the Federal Geographic Data Committee (FGDC) standard for geospatial metadata documentation.

Data Deletion Schedule

Upon completion/expiration of a grant or contract, all geospatial data produced during the life of the project should be examined by the principal investigator. Geospatial data should be inventoried and sorted into one of two general categories (retain or delete). If a geospatial dataset satisfies the retention determination described above, that data must be retained for permanent archiving. If it does not satisfy the criteria, deletion of the geospatial dataset is at the discretion of the investigator. If the principal investigator is unsure of the correct status for a geospatial data set, he/she should contact ISU's GIS Director for clarification.

Data Storage Guidelines

Preliminary backup

Initially all geospatial data satisfying the criteria set forth above will be copied to the GIS TreC's spatial library server or other server meeting the following hardware requirements:

- Hardware RAID 5 fault tolerance
- All hard drives must be hot-swappable
- Dual redundant power supplies

These hardware requirements are critical to insure safe and secure storage of geospatial data sets to aid in the prevention of loss due to corruption.

Permanent Archive

All geospatial data meeting the criteria set forth above for permanent archiving will be stored on the GIS TReC's off-site server or other off-site server meeting the following requirements.

- Hardware RAID 5 fault tolerance
- All hard drives must be hot-swappable
- Dual redundant power supplies

Alternatively, external solid state drives may be used and should also be stored off-site once data has been copied and verified on the drive.

A viable off-site location is an important consideration and should be a site that is owned/maintained by Idaho State University and physically disconnected from the source of the original data (e.g., the GIS TReC in Graveley Hall). Ideally, this location will be climate controlled and located above grade to minimize potential flooding disasters.

A permanent archive of geospatial data stored in preliminary backup will be created at intervals not to exceed 60 days. Data should not be bundled or compressed (e.g., ZIP or TAR compression) and the media used will conform to industry standard media types. Important considerations regarding archive media are:

- Sensitivity of media: Will problems be encountered due to media sensitivity to heat, humidity, and light? This consideration is perhaps most important even though archive media should be stored in an environmentally controlled facility.
- Assurance of media: Is the data stored on the media in a format that will be readable in the future? This consideration is of secondary importance as long as the selected media conforms to current ISO and/or ANSI standards. Most available media meets this criterion.
- Longevity of media: How long will the media be viable? This concern is perhaps least important, relative to those listed above.

The GsCC recommends the use of live, real-time server based archiving solutions employing a hardware RAID 5 fault tolerance solution with adequate power monitoring and uninterruptable power supply (UPS) to allow for a safe system shutdown in the event of prolonged power outage . The above server should be physically secured and use strong authentication to better ensure the integrity and security of the archived data. In accordance with guidelines set forth by the American Society of Heating, Refrigerating, and Air-Conditioning Engineers (ASHRAE) the physical environment within which the server operates should be maintained with ambient air temperature between 20° C (68° F) and 25° C (77° F) and relative humidity levels between 40-55% (ASHRAE TC9.9).

In the event that archived data need be retrieved, the retrieval process shall be completed within 3-business days of receipt of the retrieval request by ISU's GIS Director.

Contingency and Data Assurance

It is the responsibility of the GIS TReC administration to periodically assess the viability of archived data by retrieving a randomly selected media and ensuring it is readable and functional by at least one current version of a GIS/RS software application. Should it be determined that an archive has failed, all archive data will be retrieved, tested, and all recoverable data transferred to new media. To eliminate or reduce this problem, data will be cycled to new media following the schedule given below:

- All geo-spatial data stored on archive quality CD or DVD will be read and duplicated to a new archive quality CD or DVD every 10 years or as deemed necessary by the GIS Director. Re-examination schedule will be clearly marked on each CD or DVD.

Should it be determined that a standard geospatial data format has become obsolete, all data stored in that format will be re-processed using legacy software and converted to a new and compliant format.

These guidelines shall be reviewed and revised by the GsCC as necessary with a review interval not to exceed five (5) years.

The following excerpt has been taken from the *Idaho State Board of Education Records Management Guide* (section 4).

Part I. **Electronic records (E-Records)**

1. *Requirements for E-Records.* Records generated and stored on computers and related systems must be monitored and assessed for value in relation to the SBoE and the Institution's record retention schedules. These records shall be inventoried just as if on paper, and have destruction or preservation timeframes established as with other record formats. Present Idaho code refers to ANSI standards for digitized paper or photographic files as the benchmark for recordkeeping. However, new standards are in review which address a broader and more flexible set of criteria for the management of electronic records. These are standards for the content of *information identifying the type and format of data in an electronic file*, or the *metadata*, so that across time these files will potentially remain more accessible regardless of platform and application software.

2. *Principles of E-Records Management.* The following principles of electronic records management should be adhered to in both operation and planning of the work of higher education institutions:

    - Computers and computer backups should be organized with attention to the life expectancy (retention) of the information being created and stored;
    - Institutions are encouraged to migrate on a periodic schedule to newer platforms, media, and systems those records designated Permanent according to the SBoE Records Retention Schedules, with authentication and quality assurance checks to ensure data and file integrity after transfer;

3. *SBoE Guide, State and Federal Requirements.* E-Records created in the normal course of official business and retained as evidence of official policies, actions, decisions or transactions are records subject to the management requirements of this Guide. Specific legislation may also affect retention requirements e.g. Federal Limitations Act (disabilities law), state Rules of Evidence, Uniform Code of Evidence (applicable in many states) and the Federal Income Tax Act, etc.
    - Records communicated in an electronic format need to be identified, managed, protected, and retained as long as they are needed to meet operational, legal, audit, research or other requirements.
    - Records needed to support program functions should be retained, managed, and accessible in existing filing systems outside the e-mail system in accordance with the appropriate departments standard practices.
    - Originators of e-records within the Institution are responsible for proper filing and retention of those e-records. Additionally initial (original) recipients of e-record from institutions and individuals outside the Institution are similarly responsible for filing such records.
    - *Disaster Recovery.* Information Technology administrators and internal control (and/or internal audit) staff are responsible for maintaining electronic record security, backup, and for disaster recovery plans for those records placed under their administration.