Bit, Byte, Kilobyte, Megabyte, Gigabyte, and Terabyte Matryoshka

Electronic Records Management at Penn State: A Matryoshka Design

Jackie R. Esposito, Penn State University Archivist and Head,
Records Management Services

Since 1991, Penn State has been regularly pursuing and supporting Records Management Initiatives under the direction of the Records Management Advisory Committee. In December 2005, a Digital Preservation Steering Committee was established to identify parameters around which electronic records preservation of University business records would proceed to develop. The Committee's final report was submitted to the Provost in December 2008.

In January 2009, the Electronic Records Software Specification Committee was charged to identify requirements for software development for ElectRAR, the University's proposed Electronic Records Archival Repository. After six months of meetings and conferences, the ElectRAR Report was submitted for funding and implementation considerations. Simultaneously, two other University initiatives were being undertaken: 1) determine XAM Storage Space Capabilities and 2) identify electronic records practices in the centralized business and student records systems (IBIS and ISIS, respectively).

All the activity regarding the strategic and tactical planning for the University's Electronic Records issues led me to attempt to find a visual characterization of where we currently are and where we are going. Hence, the Matryoshka dolls design concept was born.

Doll # 1 – Records Management Advisory Committee: Organizational Chart
Doll # 2 – Digital Preservation Steering Committee: Recommendations
Doll # 3 – Data Classification Schema: Categories
Doll # 4 – IBIS and ISIS Electronic Records Practices Review: Executive Summary
Doll # 5 – XAM Storage Space Test and Recommendations
Doll # 6 – ElectRAR - Electronic Records Archival Repository: Flow Charts

Each doll can stand alone as a complete, holistic project. But ideally each project builds on the good work of its predecessor and creates, at the same time, its own structure.

Matryoshka Dolls (Russian Nesting Dolls)

A matryoshka doll is also known as a Russian nested doll and is a set of dolls of decreasing sizes placed one inside the other. The number of nested figures is usually five or more. The shape is mostly cylindrical, rounded at the top of the head and tapered towards the bottom. The first Russian nested doll set was carved by Vasiliy Zvezdochkin from a design by Sergei Maliutin, who was a folk crafts painter in the Abramtsevo estate of the Russian industrialist and patron of the arts Savva Mamontov. Maluitin's doll set consisted of eight dolls. The design was inspired by a set of Japanese wooden dolls representing Shichi-fuku-jin, the Seven Gods of Fortune.

Matryoshkas are also used metaphorically, as a design paradigm, known as the "matryoshka principle." The thesis denotes a recognizable relationship of "similar object-within-similar object" that appears in the design of many layering paradigms.

Doll #1 – Records Management Advisory Committee

Mission Statement – The University Archives & Records Management Advisory Committee (RMAC) has been established (1989) to oversee the university's records management program. The committee is responsible, in conjunction with the University Archivist and Records Manager, for developing, implementing, and maintaining a university records management program. The committee will

a) review records appraisal recommendations;
b) approval both general and specific records retention and disposition schedules;
c) approve the disposition of and recommended means of protecting vital university records;
d) establish policies relating to the storage and disposition of semi-active and inactive records; and
e) Develop and oversee University records policies and procedures.

Membership – RMAC will consist of representatives from the following University offices:

a) President's Office
b) Provost's Office
c) Registrar's Office
d) Human Resources
e) Computer and Information Systems (Information Technology Services)
f) Systems and Procedures
g) Risk Management
h) Financial Auditing
i) Physical Plant
j) Research Grants and Contracts
k) University Archivist
l) Records Management
m) Business Services (Representative from Inactive Records Center Operations)

Doll # 2 – Digital Preservation Steering Committee Recommendations

The activities outlined in this report will extend over a number of years and are dependent upon support from the University administration and availability of resources for implementation. Therefore, we recommend that Information Technology Services and the University Records Management Advisory Committee jointly prepare an annual progress report that is to be submitted to the Provost and Senior Vice President for Finance and Business. A goal should be to have all recommendations implemented within a three-year timetable, by December 2011.

A special task force should be appointed in January 2012 by the Provost and the Senior Vice President for Finance and Business with the task of verifying that all systems and policies are in place, and to identify any additional issues that need attention.

Issues that should be tracked for progress include:

- Monitor AD35 for future revisions as needed.
- Evaluate and revise as necessary the uniform rules for e-discovery response protocol.
- Complete and publish Data Classification Schema and the Security Standards Matrix.
- Collaborate with University Faculty Senate Office, AIS, and DLT to establish security/access protocols for archived curriculum data (CSCS) and University business records and to develop an e-record repository for permanent archiving.
- Continue the review and revisions to central systems for centralized e-preservation, to include AIS, University Libraries, etc.
- Develop “best practices” protocols and security requirements for data in other administrative and academic offices in the distributed computing environment at Penn State.
- Refine system-wide protocol for Records Liaisons.
- Establish procedures for archiving University websites once per semester.
- Employ standard business decision workflows in conjunction with implementation of AD63.

Doll # 3 – Data Classification Schema Categories

According to University Policy AD23, "…the integrity of institutional data (including Computerized Institutional Data) must be assured. All institutional data must be protected from unauthorized modification, destruction or disclosure, whether accidental or intentional". Accordingly the following classifications are established, geared to the level of protection that must exist for computers or networks that process, transmit, store or otherwise handle the various classifications. It should be noted that computer and network resources themselves have value and can be misappropriated to use the storage, processing, bandwidth and connectivity inherent to attachment to a University network. Therefore, even machines that only handle Public information must meet a minimum standard of care to be allowed to connect.

Data Classifications:

Public – Information is intended for distribution to the general public, both internal and external to the University. Release of the data either intentional or inadvertent would have no or minimal damage to the institution in any dimension.

Internal/Controlled – Information is generally intended for distribution within Penn State only, generally to defined subsets of the user population. Release of the data has the potential to create moderate damage to the institution. (Such damage may be legal, academic (loss or alteration of intellectual property), financial, or intangible (loss of reputation).

Restricted - Data which the University has a legal, regulatory or contractual obligation to protect and for which access must be strictly and individually controlled and logged. The release of such data has the potential to create major damage to the institution. (Such damage may be legal, academic (loss or alteration of intellectual property), financial, or intangible (loss of reputation). Examples of data in this category include Social Security Numbers and Personally Identifiable Health information.

Additionally, there are data whose security is mandated by the originator and for which security must be in accord with specific restrictions associated with the originator's permitting the University or unit to use the data in whole or in part. Certain research grants fall into this category, as do certain teaching databases that contain proprietary information associated with a business or business type. Additionally, the credit card industry has issued security standards for all systems and networks that handle credit card data. All units in the University that are

involved in credit card transactions must comply with the standards. The credit card industry standards are termed the "PCI-DSS" standards. The most current version of the standards are available at:

https://www.pcisecuritystandards.org/

Additional dimensions of system and network operation that must be considered in the security architecture of any unit are integrity (or accuracy) and availability. Even data that is fully public may need additional security restrictions if it must be available without fail twenty-four hours per day.

Integrity/Accuracy - Data is maintained in the same state end-to-end. In other words, data arrives at its destination and intended application or use in exactly the same state as when it was entered. No intentional or inadvertent modification of the data may ensue in transit.

Availability – Data, systems and networks are fully operational when needed for their intended use. Some systems may have very high availability requirements (for example, no down time at all is acceptable), while for others, down time may be less critical to the operational mission. The security of systems, networks, operating systems and applications should be commensurate with the overall availability requirements.

Examples in Each Data Category:

Public:

Public data may include but is not limited to information such as:

- Campus Maps
- Directory information (where no Confidentiality Hold applies)
- Email addresses of individuals (not bulk listings of all entries data mined from central services)
- News stories (subject to copyright restrictions)

Internal/Controlled:

Internal/Controlled data may include but is not limited to information such as:

- Library Collections limited to Penn State use only
- Bulk email address listings containing all members of a major population (e.g., all students, all faculty/staff)
- Class rosters
- Employment applications unless restricted information is included

Restricted:

Restricted data may include but is not limited to information such as:

- Social Security Numbers
- Drivers's License numbers
- Personally Identifiable Health Information (PHI)

- Salary and tax information related to individuals
- Details of University Budgets
- Tenure or promotion information
- Staff employee review information
- Password or other system access control information (to include biometric identification parameters)
- Human Subject Information (May have additional security requirements as identified by the originator or the Institutional Review Board)
- Non-directory information, to include photographs of individuals unless permission has been obtained for their use
- Workman's Compensation or Disability Claims
- Employee background check information
- Admission and financial aid information
- Bursar bills that are personally identifiable
- Personally identifiable grade or transcript information
- Donor information
- Security settings or details of security configurations (e.g., detailed firewall rule sets)
- Information to/from University Legal Counsel unless otherwise designated
- Ethnicity data other than aggregate statistics
- Disability status other than aggregate statistics

Doll # 4 – IBIS and ISIS Electronic Records Practice Review Executive Summary

Business Need and Project Goals/Outcome

In December 2005, the Digital Preservation Steering Committee was charged by Executive Vice President and Provost Rodney Erickson and Senior Vice President for Finance & Business Gary Schultz to develop policy language and technology system procedures for electronic university records, including:

- Retention and use of email
- Digital document creation and digital archiving
- Electronic business decisions: on-line access and procedures
- Retention and archiving of "born digital" publication and production materials

The Digital Preservation Steering Committee issued an Interim Report in August 2007, outlining the progress that has been made, including revisions to Policy AD35 - University Archives and Records Management mandating retention of "born digital records", procedures for e-Discovery Compliance and Information Privacy Protection, and the establishment of a network of Records Liaisons within the University.

This project will concentrate on the identification, and related retention, disposition and destruction of University Records, specifically, born digital records. As a starting point, AIS completed an initial assessment (Attachment I) of the magnitude of the digital preservation initiative for the areas of AIS responsibility – primarily the central repository of born digital records that comprise the Business Information System (IBIS) and the Student Information System (ISIS).

AIS will provide leadership to continue the communication and educational efforts started by the Digital Preservation Steering Committee. We anticipate that this will be a multi-year process, with coordination and input required from multiple partner offices. Partner Offices include, but are not limited to: the Office of Human Resources (OHR), University Budget Office

(UBO), Corporate Controller, Registrar's Office, Bursar's Office, Admissions, Housing, Graduate School and the Office of Student Aid. In order to manage a project of this size, the project will be broken down into three phases.

- Phase I is the process of engaging partner offices to gather information and prepare a report which identifies the data sources within the University community covered by Policy AD35 University Records Management Policy. This includes systems maintained centrally, as well as systems maintained by partner offices. The target date for completion of the Phase I report is December 2009.

  In Phase I, AIS will work with partner offices to identify those University Records (born digital) which are housed within the partner organizations. This will include both centralized systems and decentralized systems. It may be possible to use data from some of the recent survey instruments to gather some of this information. Recent surveys include the Risk Assessment, IPAS, and Data Center Survey by HP. Data dependencies will be identified, where applicable. Another element of Phase I will be to determine if Retention Schedules exist and are current. In some cases, new Retention Schedules may be required.

- Phase II will determine alternative methods to backup, archive or destroy data as directed by Policy AD35 University Archives and Records Management. If new Retention Schedules are needed, these will be developed in Phase II. Phase II will also prepare a report to evaluate the alternative methods and recommend best practices for retention, disposition and destruction of University Records (born digital). The target date for completion of the Phase II report is December 2010, which allows time for presentations and discussion of the alternative solutions in the early months of 2011. This timetable allows for budget decisions and funding for Phase III in fiscal 2011/12.

- Phase III will be implementation of the selected solution for backup, archival, or destruction of University Records (born digital). Implementation includes an implementation plan and oversight of the implementation process. Implementation is targeted to begin in the Fall 2011 after funds are allocated.

Doll # 5 – XAM Storage Space Test and Recommendations

Goals

The goals of the project are to:

1. Test capability of discovering objects originating from diverse repositories and residing on one or more XAM-aware storage disks

2. Test policy-driven storage management capabilities, such as automatic assignment of digital objects to tiered storage devices

3. Test enhanced storage system functionalities, such as availability, resistance to data loss, data de-duplication, and fault tolerance

4. Develop storage management metadata attributes

5. Promote integration of the XAM standard with various vendor-neutral storage devices

6. Support integration with existing repository software using a storage services REST API, or by means of new application development using XAM libraries

Project Outcomes

1. Demonstrated the capability to store objects and metadata to XAM and non-XAM storage arrays, and to parse metadata into and out of the storage cloud for UI dissemination and routing of objects

2. Implemented a REST API framework to simplify and enhance the adoption of XAM as a standard for saving fixed content with metadata to a storage cloud

3. Demonstrated policy-based retention controls using metadata

4. Demonstrated basic search capabilities

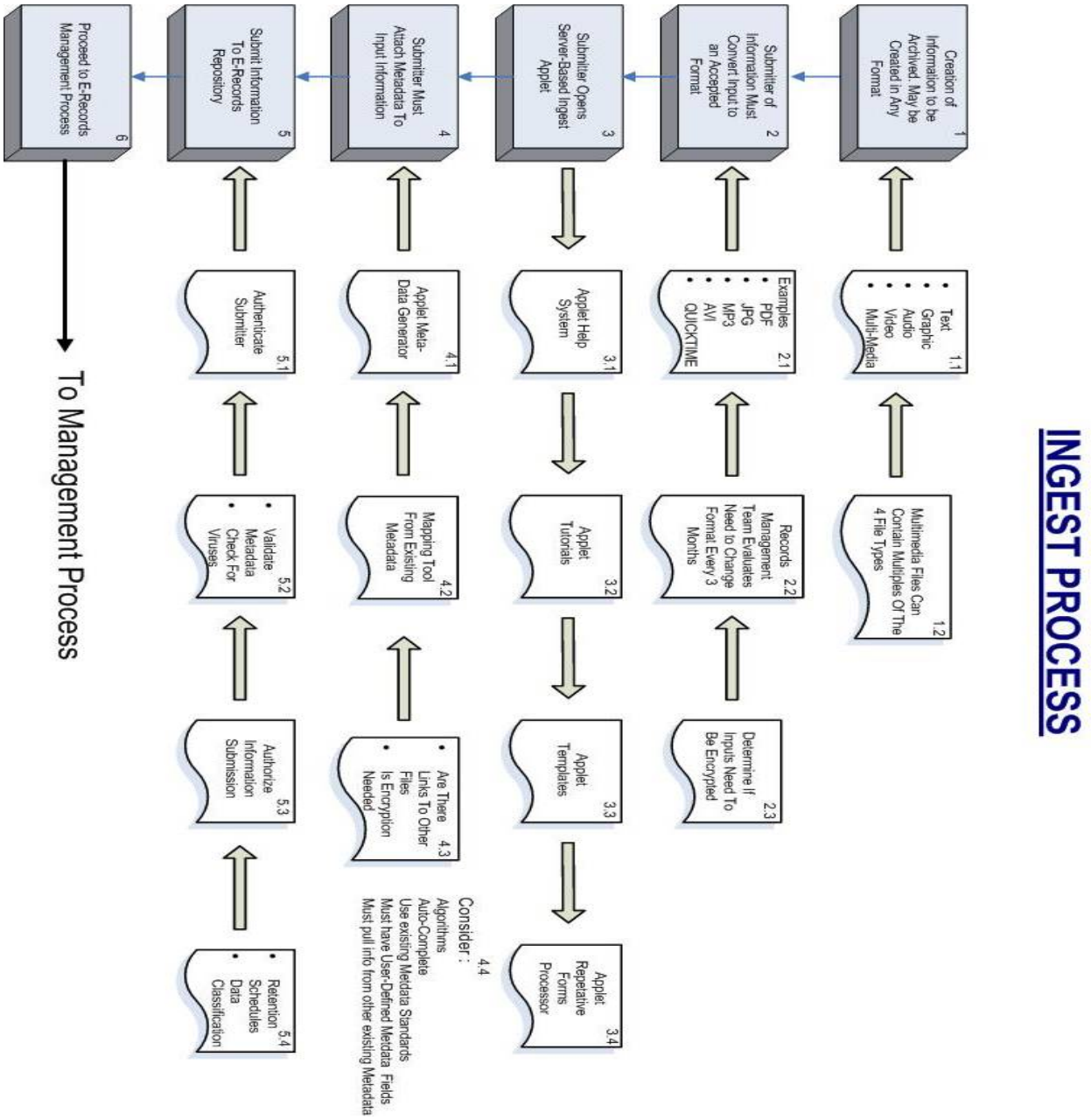
5. Developed and implemented a working code set running on the Storage Services Gateway

6. Established a stable development environment to serve as the initial Java development effort. Ultimately changes may occur to this environment over time as the project developers become more familiar with tools and processes to enhance and improve their productivity

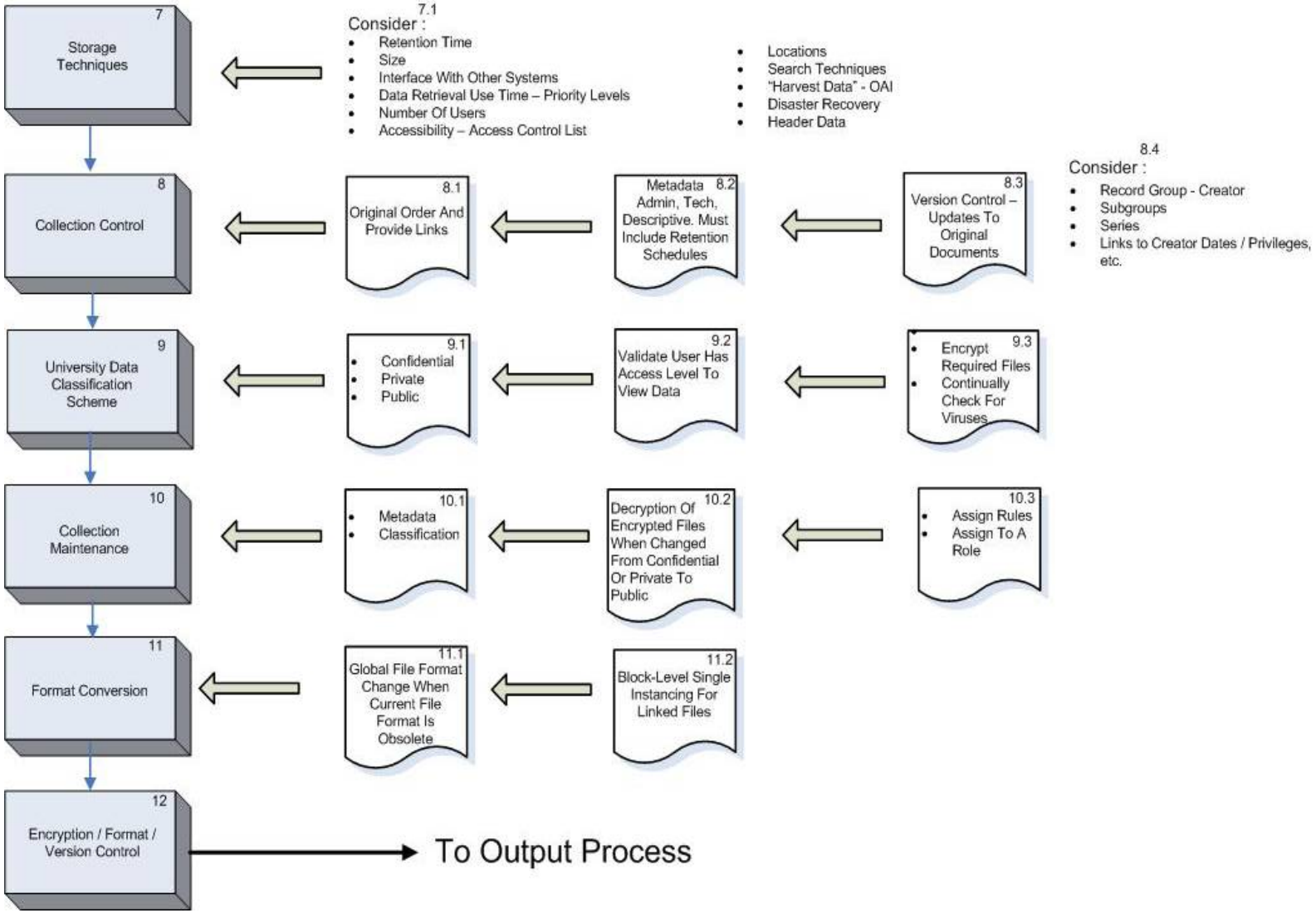
7. Implemented a unique ID for each object and demonstrated binding and non-binding effects (version control) for saved content

8. Worked successfully with storage industry representatives to promote the adoption of XAM across all types of storage device

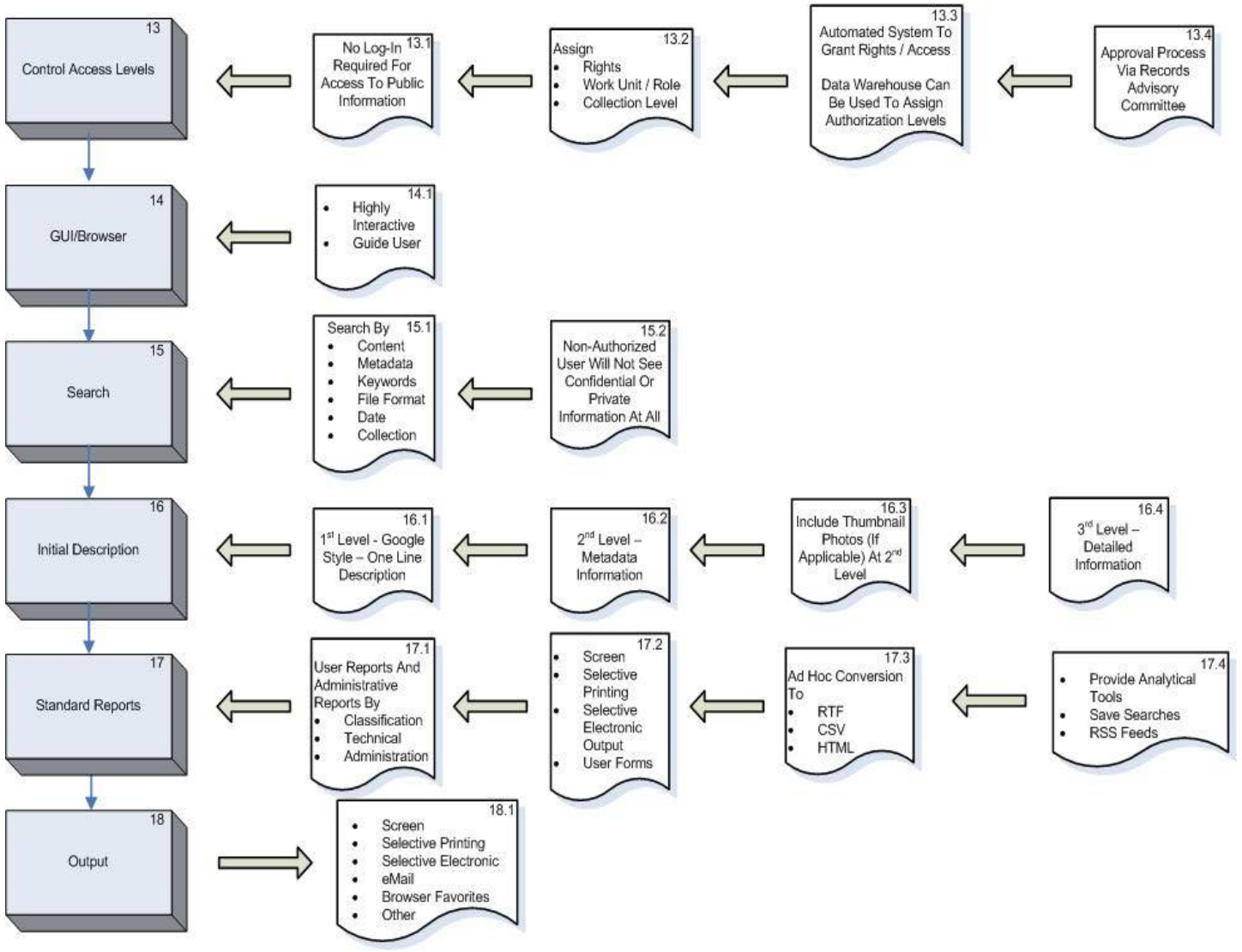
Doll #6 – ElectRAR: Electronic Records Archival Repository Flow Charts



INFORMATION MANAGEMENT PROCESS



OUTPUT PROCESS



SUMMARY

UNIVERSITY E-RECORDS ARCHIVAL REPOSITORY [ElectRAR]

All Penn State University employees must ensure that electronic records are maintained so they are readily available for appropriate use, and that established records management procedures, including retention, disposition and destruction schedules, can be carried out. ElectRAR will provide a centralized repository for these services as outlined in AD-35, the University's Archives and Records Management policy [Appendix 1A].

ElectRAR, a digital surrogate of the University Archives, will insure the capability of reconstruction of the events within the University during a specific period of time to document their historical, fiscal, administrative, and evidential value. The ElectRAR will serve three (3) primary purposes relative to University records:

1) Actively maintain and manage born digital records in a usable information structure for a recommended period of time, possibly as long as seventy-five (75) years, as mandated by University retention schedules and legal mandates. This repository will conform to the three (3) major criteria for long-term digital preservation: **authenticity, reliability, and integrity**.

2) Provide navigational guidance via a user interface (GUI) for specified access to stand-alone University repositories, designated by the Records Management Advisory Committee (RMAC)[Appendix 1E], for faculty staff, students, and researchers. The repository will become part of the University's cyber infrastructure and will function as an information management tool.

3) Provide best ingest, management and output practices for the guidance of other stand-alone University e-repositories.

This Software Requirements Specification provides the structure by which the ElectRAR can operate, manage, and preserve University electronic records for the next century. This summary includes Process Flow diagrams for:

Ingest –

The process by which the contributor is guided in entering an object (documents, photographs, audio, video or multimedia presentation) into the ElectRAR and create the associated metadata links to the objects and files.

Information Management –

The University's E-Records Archival Repository (ElectRAR) is required to manage all objects and files for their lifetime which may in some cases exceeds seventy-five (75) years. This process details the steps required to maintain authentic and reliable records with the original integrity as required by federal, state or local law and Penn State policies and procedures.

Output –

The process by which a user can access the information contained in the ElectRAR. The output must be fast, intuitive, complete and able to assist a user in developing a successful search and retrieval operation. It will also comply with all regulations for E-Discovery.

ElectRAR will serve as a valuable digital information asset for the entire University community.