Detailed Work Plan

Year One: Develop Preservation Framework

1. Profile Current and Future Media Art Researchers and Their Needs
2. Collection Analysis and Selection of Classes
3. Identify Significant Properties and Build Out Digital Object Model / Metadata Profile
4. Build SIP Requirements Based on Findings: SIP Definition and Creation for Classes of Works

Year One Project Progress Report

Year Two: Implementation into Repository Environment

1. Automated Processing and Ingest of SIPs into CUL Archival Repository
2. Test Pull of Dissemination Information Packages (DIPs)
3. Testing Against Full Range of Assets within Selected Classes
4. Definition and Creation of Access Versions
5. White Paper Composition

CUL anticipates the following generalizable outcomes from this project:

Technical analysis of the Goldsen Archive's interactive digital holdings and documented methodology, which will be highly instructive for comparable collections

Generalizable user profiles for new media art (with access version);

A viable object data model for complex digital objects with associated metadata profile (Metadata Encoding and Transmission Standard (METS) and/or Resource Description Framework (RDF);

SIP structure to support long-term preservation of objects, provisioning for future emulation development, with associated automated workflow for SIP generation and ingest, tested and vetted by CUL's ingest process; and

A full report of this process, including financial and time expenses, so that other institutions might gauge the feasibility of a comparable project with their own complex media collections.

Year One: Develop the Preservation Framework

1) Profile Current and Future Media Art Researchers and Their Needs

Our first goal is to determine use scenarios for different kinds of media art researchers. By building out use-case scenarios, we will develop a better sense of various users' requirements for access to interactive digital assets.

For example, an arts educator might have different needs than a professional artist or art student. The needs of one might be adequately, if imperfectly, satisfied with a screen shot of an interactive artwork; another might require more information about the work's interactive properties. A narrative description of interaction might be enough for an arts researcher, whereas a software historian might require information about file types, information architectures, original storage devices, pre-migration platforms, or code.

Because of the technological complexity of preserving access to these assets, CUL may not be able to fully render these pieces through time with existing technology, but we may still be able to satisfy baseline requirements for a subset of users. We need to understand the baseline access needs of current users — and anticipate the needs of future users — to build out a feasible preservation methodology.

In consultation with the media art archivists, educators, and practitioners on our advisory board, we will develop a questionnaire about visitors, researchers, and research inquiries to media art archives in this country. We will submit this questionnaire to archivists, academics, and practitioners via listservs and direct inquires, and analyze the results to develop a better profile of how people use media artworks in research repositories. Equally important, we will conduct a number of one-on-one interviews with researchers to further our understanding of their needs. Our findings from this initial research phase will guide the rest of our project.

Activities

1. Develop generalizable survey for determining varying needs of researchers working with new media objects;
2. Working with Advisory Board, disseminate survey to key set of researchers;
3. Interview a subset of users one-on-one
4. Analyze results to determine generalizable user profiles with associated access requirements.

Deliverables

1. User profiles with associated requirements for long-term access of new media art.

2) Collection Analysis and Selection of Classes

To develop our metadata and SIP requirements and subsequent automated ingest of complex media, we need better technical data about the contents of the collection itself. This need demands a better way of evaluating the nature and risk of the various pieces within the collection. Comparable assessments have been undertaken with collections of complex digital objects, such as the Preserving Virtual Worlds project[1|#_ftn1]; however, no test bed for forensic assessment has been as broad, rich, complex, or wide-ranging as the interactive holdings of the Goldsen Archive.

We will capture file formats, hierarchical structure and relationships, hardware and software requirements (including operating systems and browser support), and other technical elements in an automated and systematic way. In consultation with the project advisors knowledgeable in digital forensics, and referencing similar projects already undertaken at the University of Maryland and Stanford University, we will evaluate the entire interactive collection. We will use this assessment to identify high-risk material based on risk of obsolescence for hardware, software, or browsers; material degradation or bit-rot; and critical dependencies such as relational and file structure contingencies. This assessment will establish the asset categories based in information architecture and technological risk-level for the next phase of the project. Ultimately, this analysis will provide a foundation for the preservation track by offering a baseline profile of the Goldsen Archive's holdings. Establishing an automated process for collection assessment would prove invaluable for comparable collections of interactive digital material.

To reach our goal of developing a comprehensive Submission Information Package (SIP) that will contain appropriate content, metadata, and documentation required to support long-term access to new media artworks throughout evolving technological landscapes, this analysis of artwork characteristics will inform the required documentation for various classes of works at the hardware, software, operating system, and file levels. The analysis of the collection will inform the development of groupings, or classification of works, which share common representation information,[2|#_ftn2] that can be used to form the initial structure of a SIP, and thus establish SIP classes.

The formation of a data model that will capture those required dependencies and structural information will begin at this phase. The data model definition will start with the identification of critical information entities that will need to be captured (processors, input/output devices, operating systems, software, libraries, file groupings), and the relationships between those entities. This phase will lay the groundwork for the later development of the detailed attributes of each of these entities.

From this assessment, we propose to select two to three distinct, but related, "classes" of material to test; for instance, a "class" might consist of a group of works created with the same software; related "classes" might represent a single software environment that functioned in CD-ROM and migrated to the web. We will, however, let our findings and the advice of our consultants guide our selection in this project phase. In making this selection, we will primarily look for categories that have: large impact---that is, exhibit information structures with potentially broad prevalence, even outside the Goldsen collections; good chance of success, and seem particularly viable for migration and potential future emulation; and scholarly value, specifically, we will seek categories that represent especially culturally significant artworks.

Activities

1. Develop framework and methodology for analysis of complex objects and identification of classes of works that share characteristics and dependencies;
2. Analyze CD-ROM and Internet Art to determine classes and groupings based on shared characteristics and dependencies;
3. Identify/develop appropriate data model for documentation of classes and representation information, beginning by reviewing existing data models;
4. Working with advisory board, select subset of classes of material to test, based on broad impact, feasibility, and scholarly value;
5. After selection by Cornell and the Advisory Board of the two or three priority classes, document representation information for each class using the data model; and
6. Revision of data model and classes as necessary based on findings from step 5.

Deliverables

1. Framework and methodology for analysis and classification;
2. XML document for collection's item-level metadata as captured in broad-stroke forensic analysis;
3. Data model for classes and representation information with accompanying documentation; and
4. Population of the data model for each class as parsed from digital forensic analysis.

3) Identify Significant Properties and Build Out Digital Object Model/Metadata Profile

We will next hone in on specific preservation requirements for works within the selected classes of material. Whereas we initially focused on a broad-stroke analysis at the collection level to determine different classes of material, this next phase will detail the necessary technical components needed to preserve and render the artworks into the future for items within those classes. We envision this as a multi-step process.

First, we must develop the framework and methodology for analysis of complex objects within each class. A key challenge will be to define the rationale for selecting a significant property_,_ and documenting these explanations in sufficient detail and quality to serve as a project deliverable. For example, within a given class of material, certain core behaviors (such as interactivity) are rendered by specific technologies. Technological pieces that enable those behaviors would qualify as significant properties, which need to be articulated, analyzed, documented, and expressed in a defined structure. Decision-making and research will be needed to identify which properties are truly significant and the implications for when those properties obsolesce. In establishing our criteria to define those significant properties, we will make provisions for future advances in emulation.

With a methodology established, significant properties will be defined for the selected classes of works with the aim of addressing the detailed attributes of the various components required to render a work, both from a purely technological standpoint, and as they relate to the work's intended behaviors, display, and functionality. This will involve a breakdown of the artworks' rendering environments, to determine what technical components are mandatory for long-term preservation and access.

Lastly, we will refine the initial data model developed to capture the significant properties for selected objects. This will define the technical and administrative metadata required for preservation of complex objects and the subsequent metadata framework. As a subset of the administrative metadata, we will capture rights management metadata — including but not limited to information about donor agreements and licensing contracts — to facilitate decision-making on use and preservation strategies. This may be included as PREMIS Preservation Metadata within the METS wrapper or built into the RDF and ontology.

We will likely build on METS and/or RDF. Rigorous analysis of the strengths and weakness of each will be conducted, addressing their suitability to effectively express complex metadata, meet local metadata requirements, and meet the needs and goals of our partners and sister collections.

METS is well-established in the library and digital preservation communities as a logical wrapper of information packages. It provides a means for expressing the structure of metadata and content files within a package, but is limited in how it can express relationships between those files, and also between the files and metadata that describe them. In contrast, RDF, and an associated family of standards, including RDF/XML, OWL, SPARQL and SKOS, provide a highly flexible means for expressing not only relationships between files, but between all concepts (i.e. metadata) about those files. While METS is widely adopted, and even locally implemented at Cornell for structural metadata during ingest into our digital archival repository, many organizations (including the Preserving Virtual Worlds initiative[3|#_ftn3]) are moving toward RDF and developing OWL ontologies for defining semantic metadata (i.e. meaningful) that can be expressed and understood by machines, providing an additional layer of value over the traditional syntactic methods of encoding metadata. Further, RDF allows for flexible and expandable description, which could aid long-term iterative expansion of preservation metadata for these objects.

Regardless of the decision, our goal is to produce shareable results. If METS is chosen, a METS profile will be developed and submitted to the Library of Congress for registration and use by other organizations with similar collections. If RDF is chosen, an OWL ontology for the expression of data about interactive media art will be developed and disseminated. Both options will likely be tested during this project, and thus both the METS profile and the OWL ontology may be disseminated.

Activities

1. Develop framework and methodology for analysis of complex objects and identification of significant properties (similar but distinct from previous methodology at collection level);
2. Perform analysis of complex objects within selected classes to identify their significant properties;
3. Continue development of data model and add documentation of complex object component parts and significant properties for selected objects;
4. Build-out METS profile and/or RDF/OWL ontology; and
5. Revision of data model and significant properties, as necessary.

Deliverables

1. Framework and methodology for analysis and identification of significant properties;
2. Revised data model, including significant properties;
3. Continued population of the data model for selected works;
4. Publishable METS profile and/or RDF data model and ontological work.

4) Build SIP Requirements Based on Findings: SIP Definition and Creation for Classes of Works

This process will involve developing packages for new media objects that contain the complex elements and metadata required to preserve them. In consideration of the Open Archival Information System (OAIS) data model, the developed SIP structure will conform to international standards and should be generalizable for other collections of new media art and complex media types.

We will build upon investigation and work already performed between CUL-Information Technology, CUL-DSPS and CUL-Metadata Services to define descriptive, administrative, and rights metadata that will be captured as part of the SIP; we will also determine how we might automate the process from existing catalog records, and augment manually as necessary.

Investigation in this phase will define the content, metadata, and packaging specifications for the SIPs as they relate to each class. We also assume that the hardware/software described by representation information will not be captured as part of the SIP (i.e., we will not address wrapping Mac OS9 in the SIP, but we will document it and its relevant properties). This process will be documented in sufficient detail and quality to serve as a project deliverable.

Activities

1. Identify required metadata beyond representation information and significant properties, such as descriptive and rights information;
2. Define content, metadata, and packaging specifications for identified classes;
3. Develop SIP creation protocol and methodology for each class; and
4. Develop SIP validation definition.

Deliverables

1. SIP specification for each selected class;
2. SIP creation protocol document for each selected class; and
3. Sample valid SIPs.

Year Two: Implement into Repository Environment and Publish Results

With a SIP structure in place for the articulation and documentation of key significant elements needed for the preservation of a complex digital object, we will begin the process of pushing assets to the CUL Archival Repository. This will be an iterative process, whereby we uncover potential problems within the preservation framework and revise accordingly. We have the following goals established for this phase of development:

1) Automated Processing and Ingest of SIPs into CUL Archival Repository

With our data model established at the class level, we will need to build out the associated workflow to package and ingest the SIPs into the digital archival repository in an efficient manner. To this end, we will investigate automation strategies such as BagIt and Dflat to streamline the process. Given the complexity of the material in terms of file structure, file type, and description, we anticipate that an individual artwork will comprise a SIP, and we will build out our workflow accordingly.

Once we have an established ingest methodology and automated the creation of SIPs, we will validate them and ensure data integrity through the use of checksums. Although much of this research will be contingent on our specific repository environment, we believe that it will be generalizable enough to provide a starting point for other institutions to do similar work. All work will be documented in sufficient detail to serve as a project deliverable.

2) Test Pull of Dissemination Information Packages (DIPs)

Ultimate success of this project relies on the ability of others outside of Cornell's current institutional infrastructure to take a DIP and use its contents. To this end, and working closely with our advisory members, we will select a partner institution knowledgeable about new media artwork and preservation infrastructure to take one of the CUL packages and validate and understand its contents.

3) Test against Full Range of Assets within Selected Classes

Testing the developed model is an integral part of any R & D effort. Following the model development, it will be tested comprehensively for the two or three defined classes. Ideally, comprehensively ingesting the classes to CUL's digital archival repository will result in the actual preservation of the assets within the collection.

4) Definition and Creation of Access Versions

In instances when rendering the original is too onerous a task or technically unfeasible, it helps to define an acceptable access copy for a given class. These may include screen grabs, screen recordings, video recordings, images, and thorough documentation of the work. When an access copy is generated, it remains to be determined whether it is incorporated into the SIP. Our efforts will be informed by the user profiles generated in the first year, in collaboration with the Advisory Board.

With an established methodology, we will create access versions for pieces in the priority classes as an initial test, which will inform the ultimate access strategy for the entire Goldsen Collection.

5) White Paper Composition

One primary deliverable will be a white paper that explains our project and its associated outcomes. All technical strategies and specifications will be articulated, as well as the factors that informed our decision-making process. We will also address a range of practical issues, including costs associated with the study, resources required to move the project into a program, lessons learned, and scholarly uses of new media art to enable us to estimate future use by scholars and students