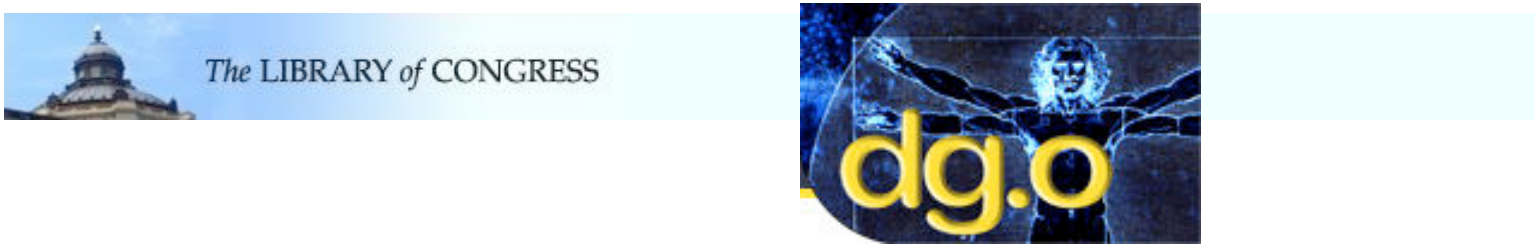


Digital Preservation Lifecycle Management

Building a demonstration prototype for the preservation of large-scale multi-media collections

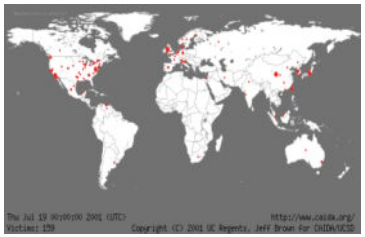
Arcot Rajasekar
San Diego Supercomputer Center,
University of California, San Diego

A Deluge of Data

- Today, data comes from everywhere
  - Scientific instruments
  - Experiments
  - Sensors and sensor nets
  - Video Streams
- And is used by everyone
  - Scientists
  - Consumers
  - Educators
  - General public
- Cyber environments must support unprecedented diversity, globalization, integration, scale, and use

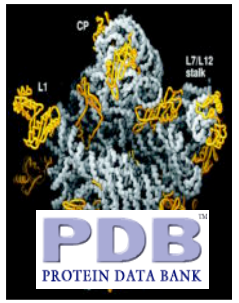
Data Management and Mining

Geosciences



Modeling and Simulation

Life Sciences



Preservation and Archiving

Astronomy

Extreme Curation – Long-term Preservation

- How can we archive, access, and preserve valued community collections for extended timeframes?

The Shoah Collection houses irreplaceable video and digital testimonies from victims of the Holocaust

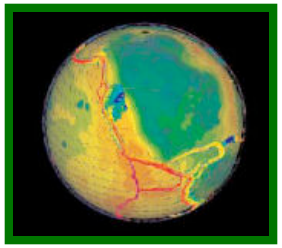
IRIS

NATIONAL SCIENCE FOUNDATION

NARA

The Library of Congress

Abraham Lincoln Presidential Library and Museum, due to open in early 2005. The site will include preservation of a broad set of materials – from paper to the presidential web site itself.



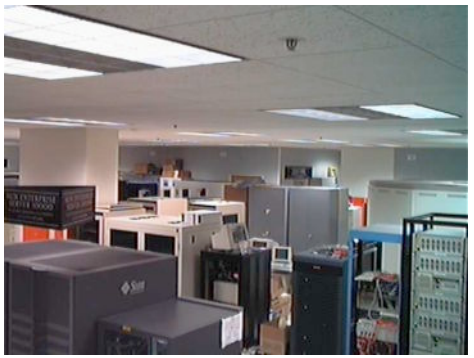
To best validate and verify long-term climate predictions, data must be maintained over 10s to 100s of years.

The 1915 film "Where the Road Divided" was printed on nitrate stock. Preservation for nitrate films must be copied to acetate "safety stock" and cannot be preserved even in controlled archival storage.

Key Challenges

- What should we preserve?
  - What materials must be "rescued"?
  - How to plan for preservation of materials by design?
  - What is the "original"?
- How should we preserve it?
  - Formats
  - Storage
  - Media
  - Facilities
  - Stewardship
  - Preservation metadata

Print media provides easy access for long periods of time but is hard to data-mine

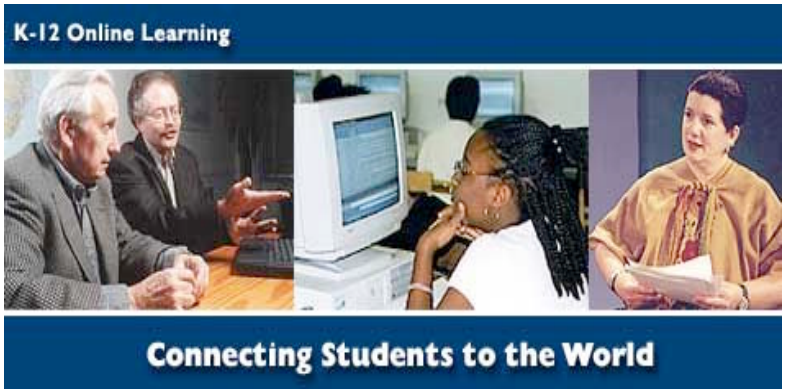


Digital media is easier to data-mine but requires management of evolution of media and resource planning over time

- Who should pay?
  - Business models for "initialization"
  - Business models for "steady state"
- Availability
  - Who should access digital materials?
  - What tools should be provided for access?
- Who's responsible?
  - Collection generators (researchers)
  - Scientific disciplines
  - Libraries
  - Government agencies

Our Proposal

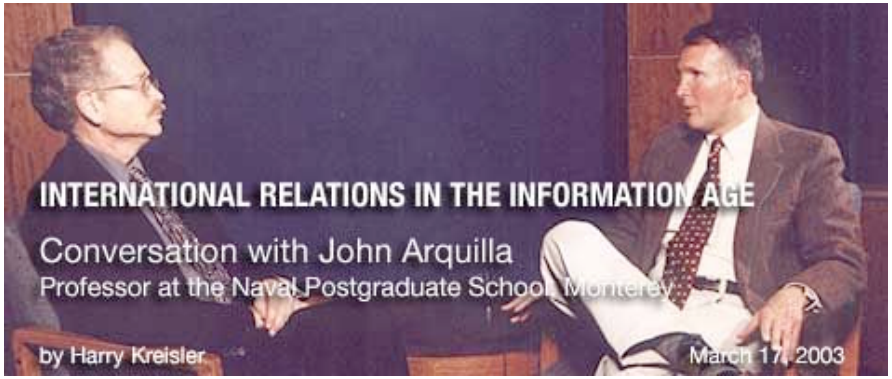
- Design and Development of a Prototype for Preserving Digital Video Collections
  - Management of Authenticity, Integrity, and Infrastructure Independence
  - Preservation Life-cycle meshing seamlessly with the content production
    - Minimal impact to production life-cycle

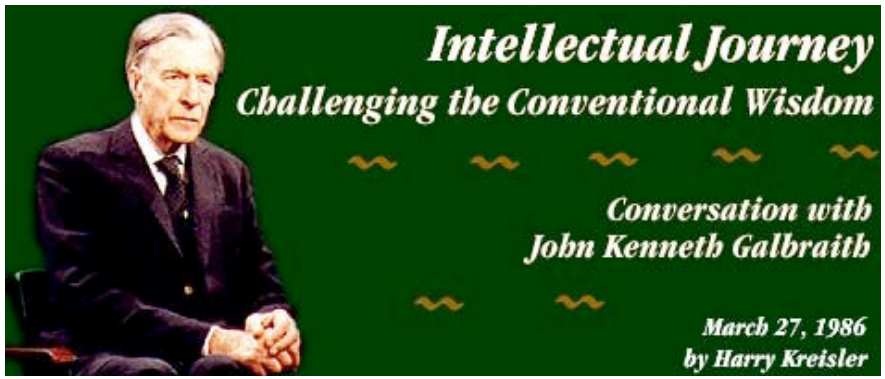


  - Workflow system that automates accession, description, organization and preservation of video and associated contents
    - Metadata definition, extraction and ingestion
    - Long-term retention and technology migration
  - At risk Collection: 'Conversation with History' video collection
    - **Video, audio, text transcripts, web-based material**
    - **Databases of administrative and descriptive metadata**
    - **Derived products**
  - Time-line: 1 year

Exemplar Collection

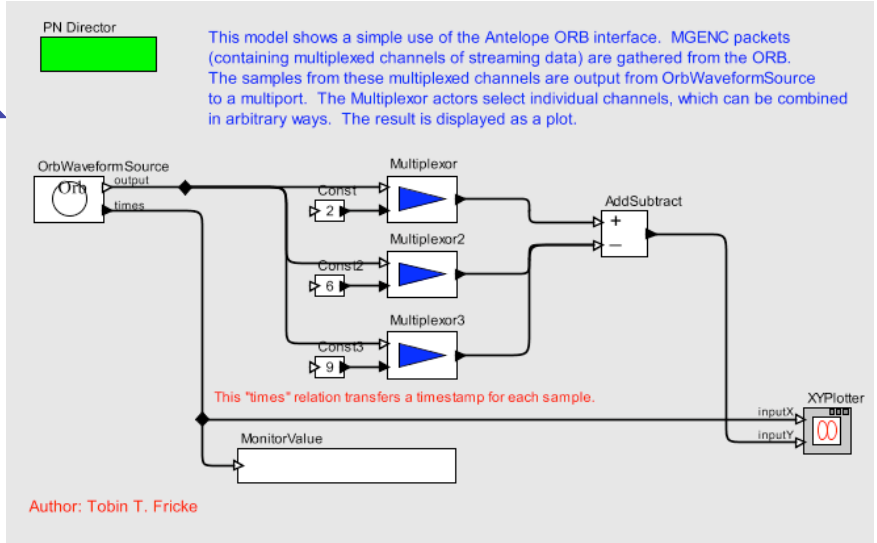
- Conversation with History - UCTV - from 1982
  - Hour-long interviews with internationally prominent individuals
  - Institute of International Affairs, UC Berkeley
  - Available in 15 million homes nationwide via UCTV
  - 40 program segments annually
  - Web-site for downloading older segments
  - Among UCTVs most accessed on-line programs
  - Programs used in educational material



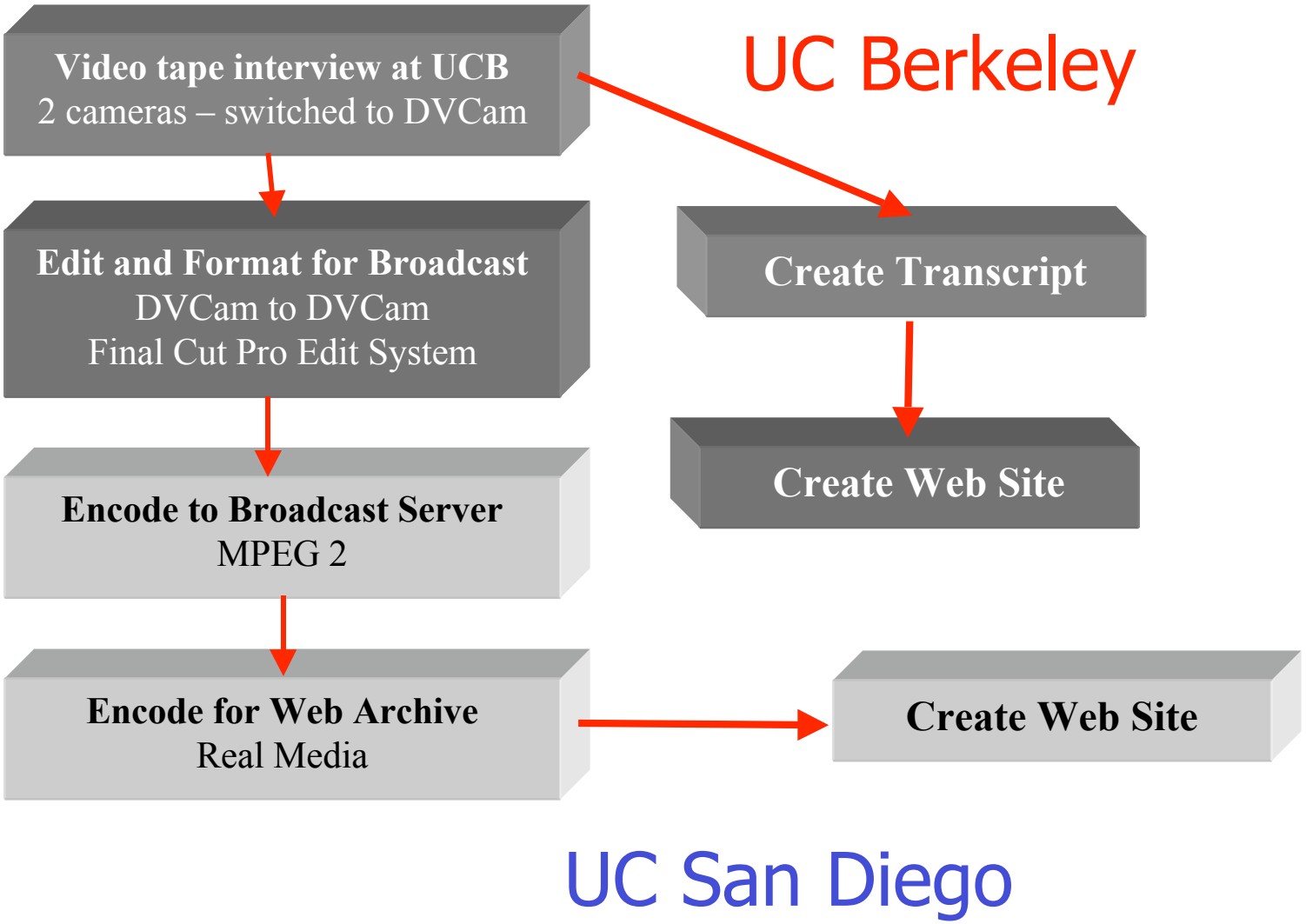


Research Plan

- Define High-Level Life-cycle processes
- Develop Models for Multi-media Collection preservation
- Define Preservation Building Blocks – AIPs
- Integration of Video Content Generation with Preservation Life-cycle
- Issues in Security, Trust & Collection Management
- Long-term Preservation Models - Chronopolis



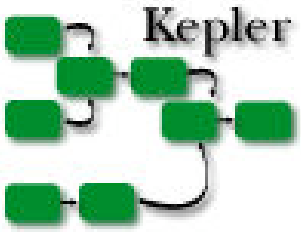
Content Production Workflow



Preservation Processes

- Generation of a Globally Unique Identifier (GUID) for each interview session
- Retrieval of the original video session
- Retrieval of each of the segments associated with a video session
- Retrieval of the transmission scripts for each video segment
- Retrieval of the material published on the Web page for each segment
- Processing of each Web page to redirect internal URLs into handles within the preservation logical name space for digital entities
- Retrieval of the rights statement for each session
- Retrieval of the header associated with each video segment
- Retrieval of the trailer associated with each video segment
- Retrieval of the administrative, structural, and descriptive metadata stored in the Filemaker Pro database
- Retrieval of the annotations stored with the Web pages
- Specification of Preservation Metadata for AIP
- Creation of AIPs for the above material
- Creation of containers for physically aggregating material for storage in the preservation environment
- Storage of containers within the preservation environment
- Specification of preservation management metadata such as access controls, storage location, and replication

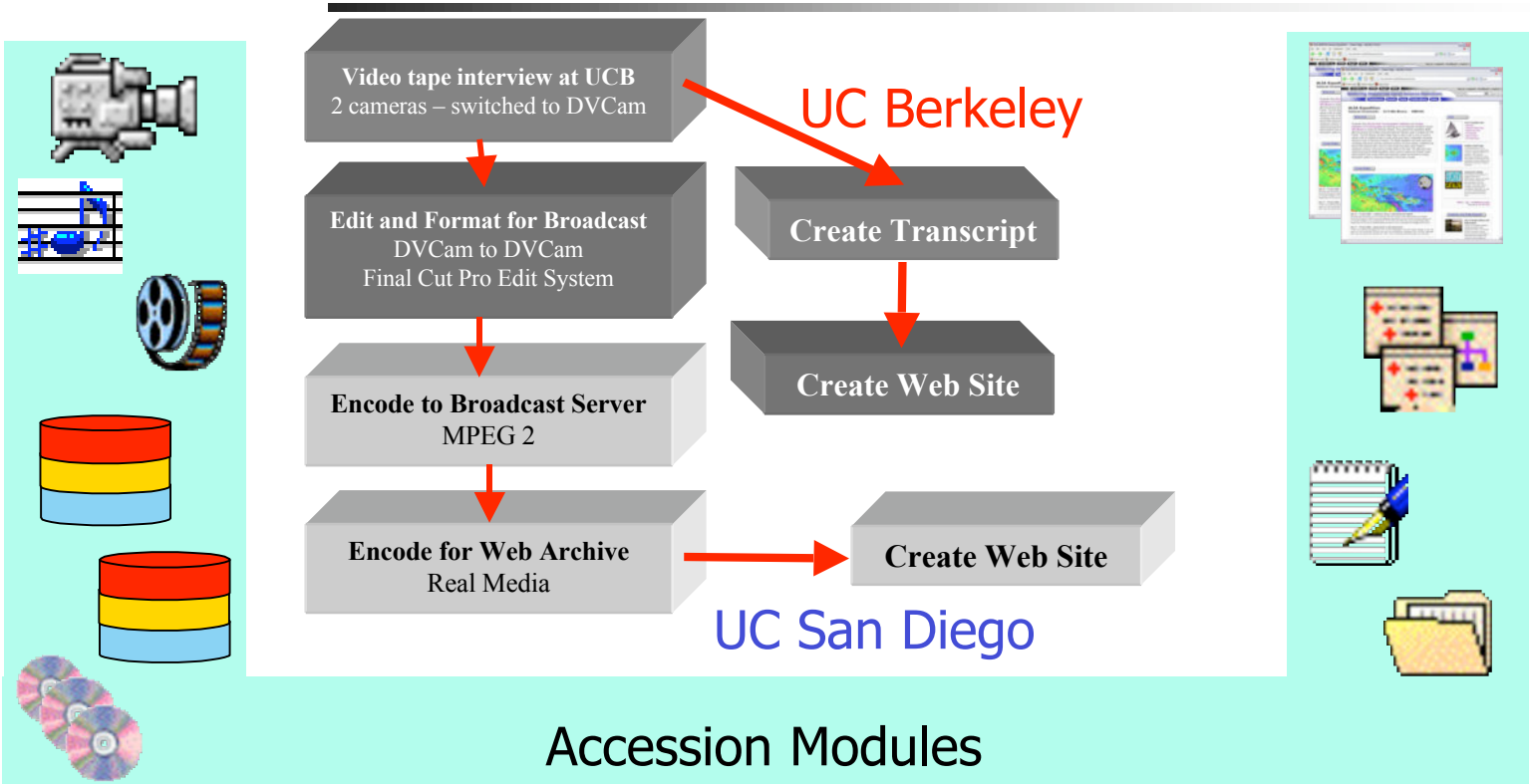
Tools



- Kepler
  - Interactive Workflow Tool
  - Used in multiple NSF projects – SEEK, GEON, etc
- Storage Resource Broker – Infrastructure Independence
  - Collection Management Tool – Organization & Repurposing
  - Uniform Access Abstraction with Access Control/Audit Trails
  - Distributed, Heterogeneous Resource Access
- Metadata Catalog – Authenticity & Integrity Maintenance
  - System, Structural, Descriptive, Preservation Metadata
  - User-defined and Discipline Specific Metadata
  - Extensible Schema Support

- Dspace
  - Accession and Discovery Tool for Documents
- Web-Preservation Software
  - Automatic Crawl of Web sites
  - Auto-translation of links

Content Generation & Preservation Workflow Integration



Accession Modules

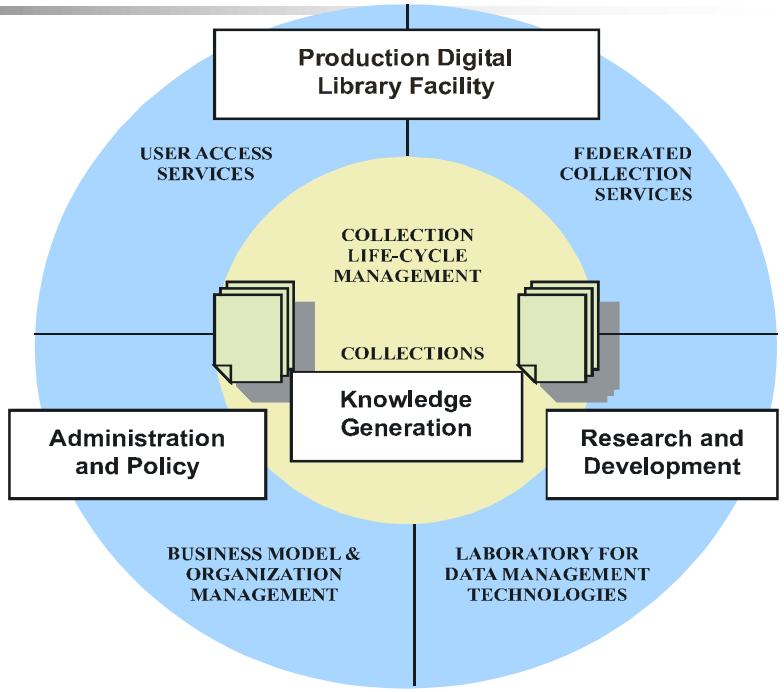
Workflow System

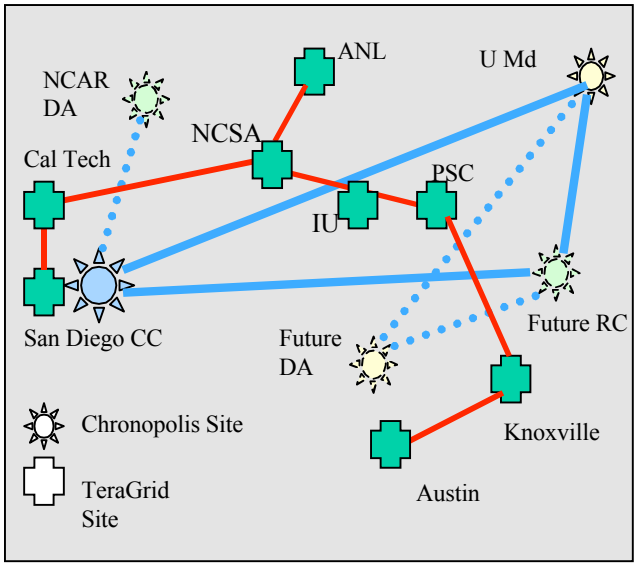
Preservation Infrastructure

Chronopolis – Long-term Preservation

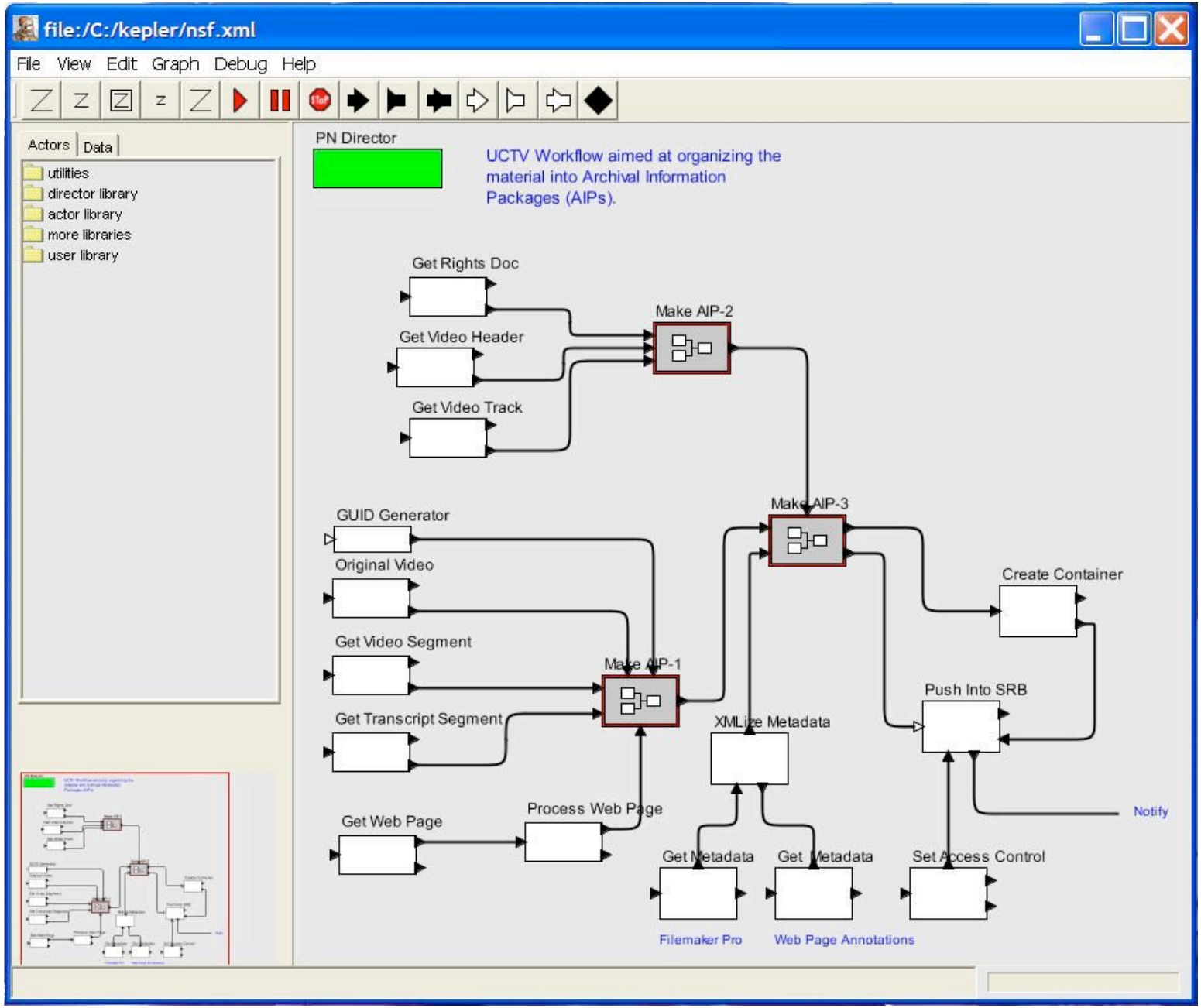
Chronopolis -- A Comprehensive Approach to Long-term Preservation

- Chronopolis integrates
  - **Collection ingestion**
  - **Access and Services**
  - **Research and development** for new functionality and adaptation to evolving technologies
  - **Collection Management** for organization,authenticity & integrity maintenance
  - **Business model, data policies, and management** issues critical to success of the infrastructure
  - **Three Sites**
    - Preservation Center (UCSD)
    - Dim Archive (UMd)
    - Dark Archive (NCAR)
  - **NSF Funding**



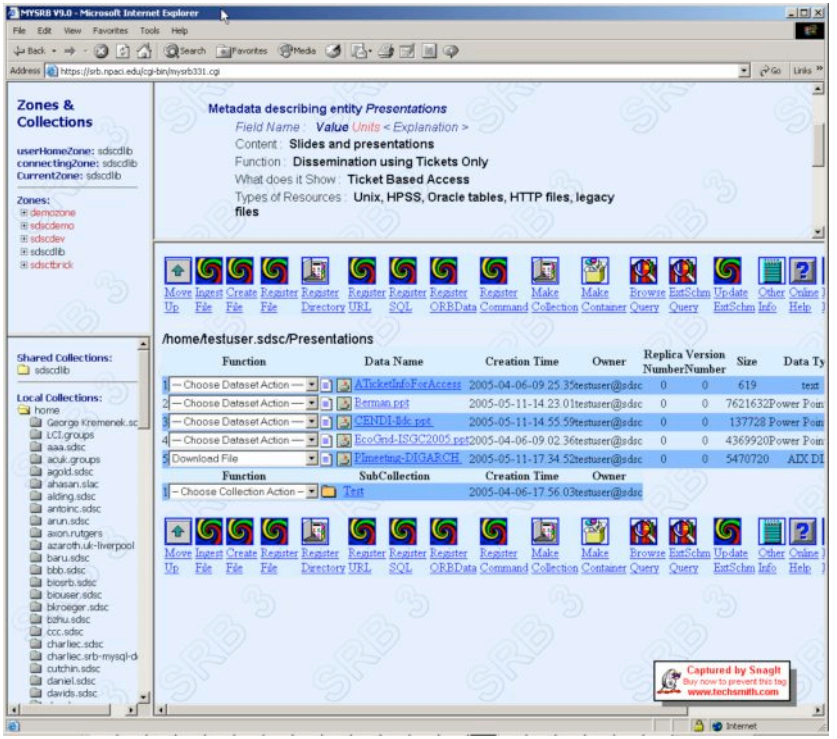


Preservation Life-cycle Management Mockup



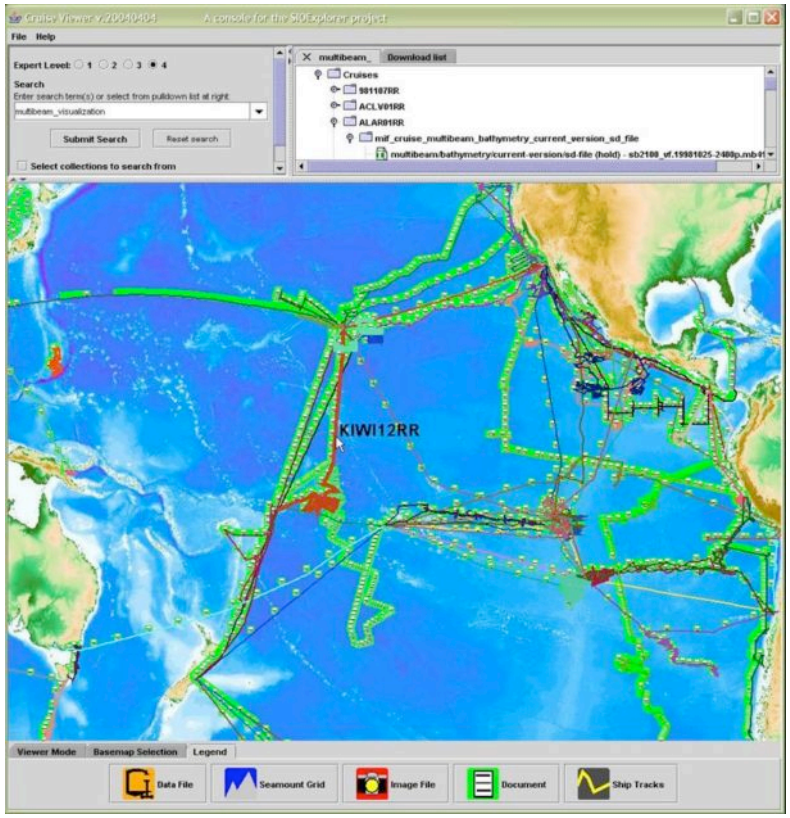
Players & Roles

- UCTV
  - Video generation and Broadcast
  - Transcription
  - Descriptive and production metadata generation
- UCSD-TV
  - Webcast generation
  - Multiple Formats
  - Descriptive metadata generation
- UCSD Libraries
  - Preservation Models and Processes
  - Description, Metadata Definition
  - Preservation metadata generation
  - Accession and Arrangement
- SDSC
  - Development of Life-cycle Management systems
  - Storage
  - Preservation
  - Access



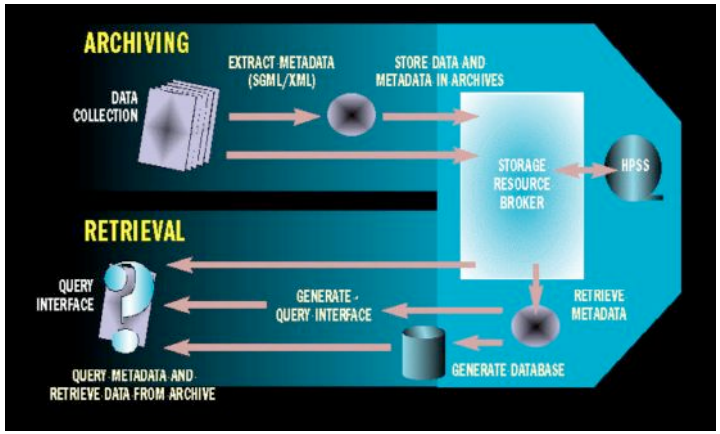
Project Management Plans

- Installation of data grid
- Access to material
  - Migration of existing videos
  - Registration of existing metadata
  - Integration with Current Workflow
- Creation of archival form
  - Validation of authenticity and integrity metadata
  - Construction of Archival Information Package
- Preservation
  - Loading into preservation collection
- Access mechanisms
  - Controlled Access, Single-sign on and audit trails
- Workshop on SRB



Research and Development

- Metadata and Data Accession
  - Human Interactions – Study of Methodology
  - Accession Module
- Design and validation of preservation templates
  - Producer-archive submission pipeline
  - Authenticity metadata extraction
  - Integrity metadata extraction



- Development of preservation workflow
  - Mapping of templates to Kepler workflow actors
  - Validation of workflows
- Extensions to the SRB data grid
  - Administrative tools for applying preservation policies

Conclusion

- Build upon a track record:
  - SDSC's collection & storage management capabilities and in developing production software
  - UCSD Library leadership role in Digital preservation for library collections
  - Strong experience of UCTV and UCSD-TV in running content generation systems
- Use Proven Technologies:
  - Kepler and workflows
  - SRB, MCAT collection management systems
  - Dspace digital content management
- We plan to deliver a prototype in one year