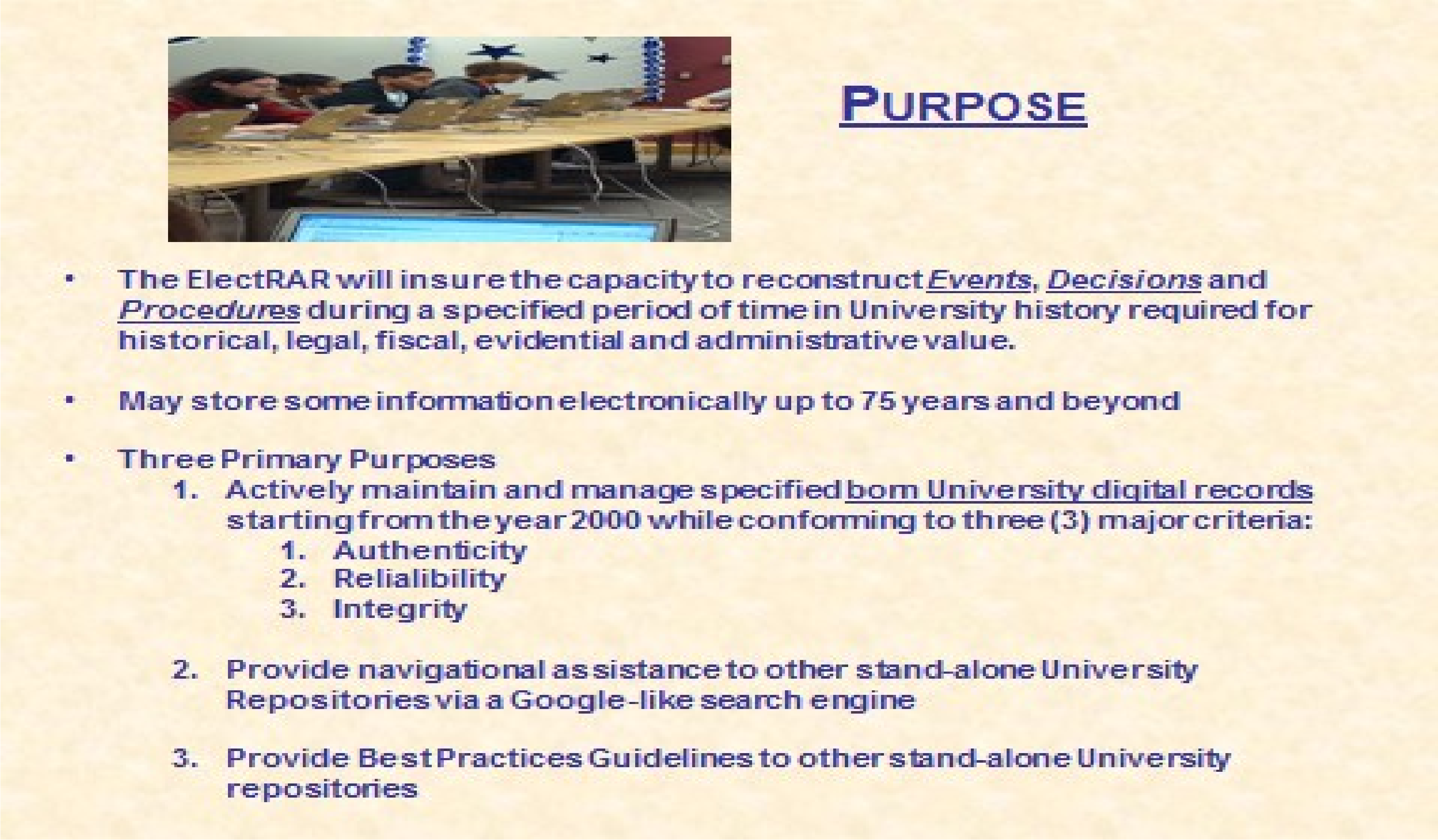
CAPS: Building Partnerships for an Electronic Records Program

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

Michelle Belden
Penn State University Access Archivist

Society of American Archivists
College and University Archivists Section
August 2011

Background - Requirements



PURPOSE

- The ElectRAR will insure the capacity to reconstruct *Events*, *Decisions* and *Procedures* during a specified period of time in University history required for historical, legal, fiscal, evidential and administrative value.

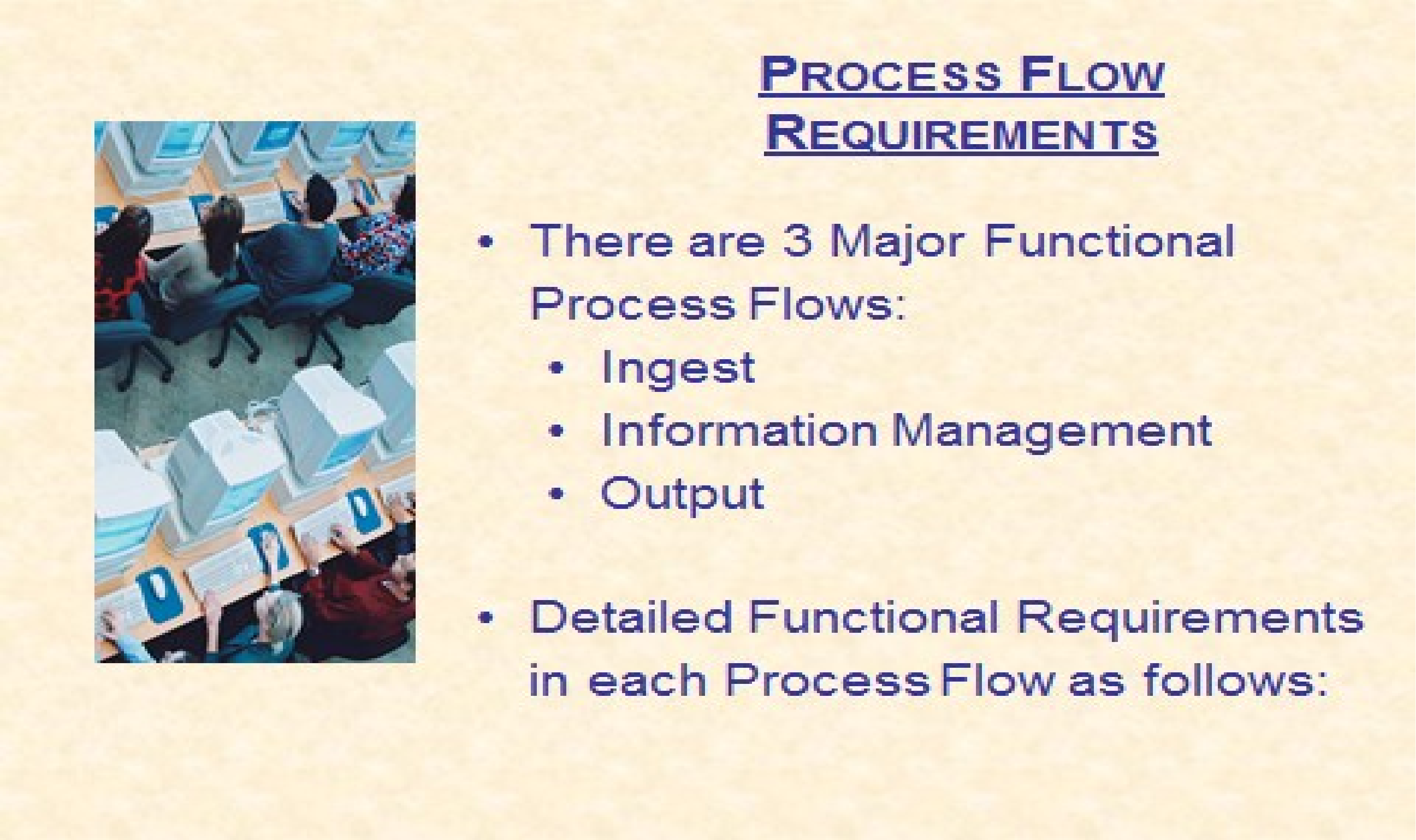
- May store some information electronically up to 75 years and beyond

- Three Primary Purposes
  1. Actively maintain and manage specified bom University digital records starting from the year 2000 while conforming to three (3) major criteria:
     1. Authenticity
     2. Relialibility
     3. Integrity

  2. Provide navigational assistance to other stand-alone University Repositories via a Google-like search engine

  3. Provide Best Practices Guidelines to other stand-alone University repositories

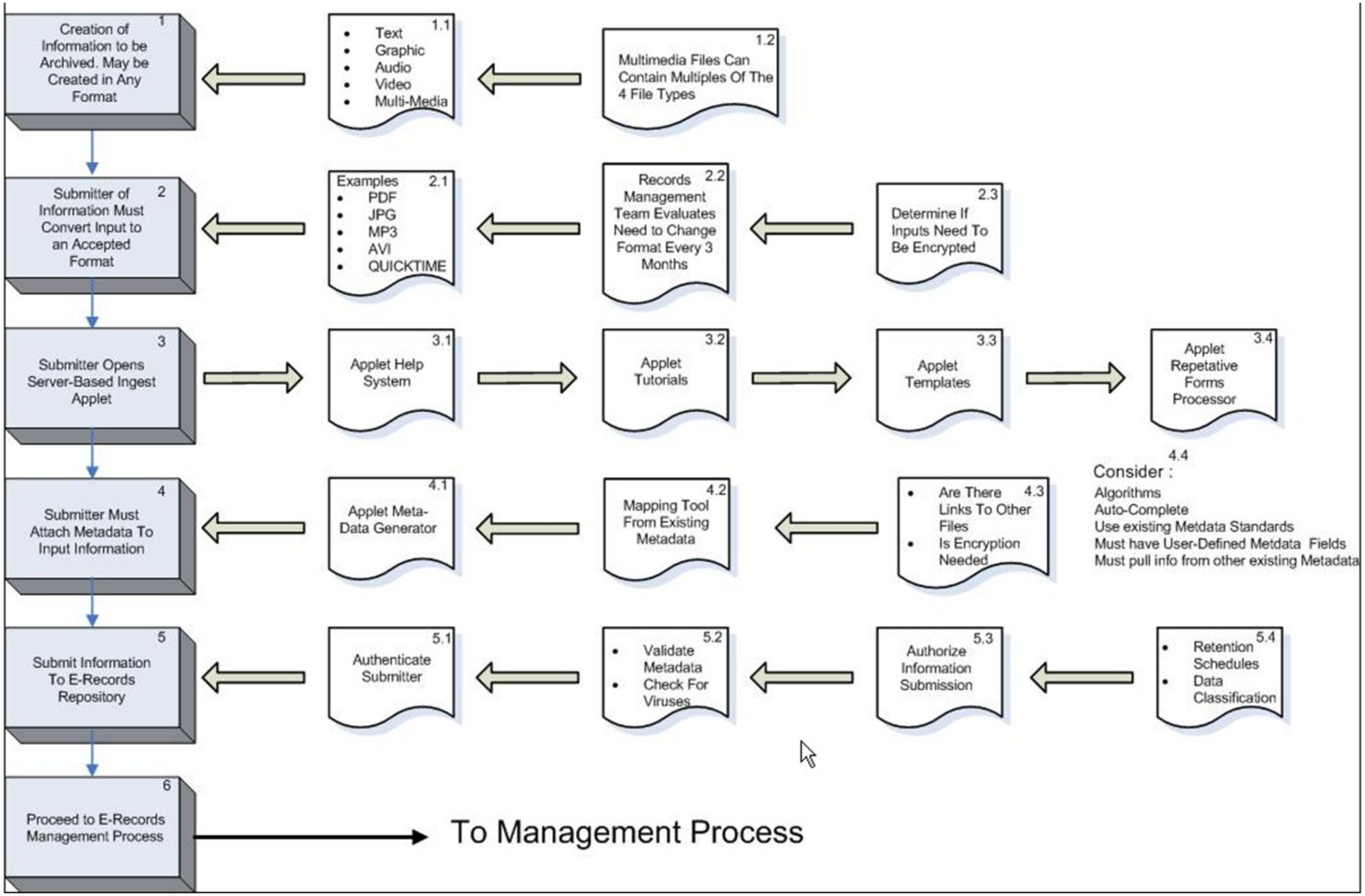
Process Flow Charts



Process Flow Requirements

- There are 3 Major Functional Process Flows:
  - Ingest
  - Information Management
  - Output
- Detailed Functional Requirements in each Process Flow as follows:

Ingest Workflow



1 Creation of Information to be Archived. May be Created in Any Format

1.1
- Text
- Graphic
- Audio
- Video
- Multi-Media

1.2 Multimedia Files Can Contain Multiples Of The 4 File Types

2 Submitter of Information Must Convert Input to an Accepted Format

2.1 Examples
- PDF
- JPG
- MP3
- AVI
- QUICKTIME

2.2 Records Management Team Evaluates Need to Change Format Every 3 Months

2.3 Determine If Inputs Need To Be Encrypted

3 Submitter Opens Server-Based Ingest Applet

3.1 Applet Help System

3.2 Applet Tutorials

3.3 Applet Templates

3.4 Applet Repetative Forms Processor

4 Submitter Must Attach Metadata To Input Information

4.1 Applet Meta-Data Generator

4.2 Mapping Tool From Existing Metadata

4.3
- Are There Links To Other Files
- Is Encryption Needed

4.4 Consider :

Algorithms
Auto-Complete
Use existing Metdata Standards
Must have User-Defined Metdata Fields
Must pull info from other existing Metadata

5 Submit Information To E-Records Repository

5.1 Authenticate Submitter

5.2
- Validate Metadata
- Check For Viruses

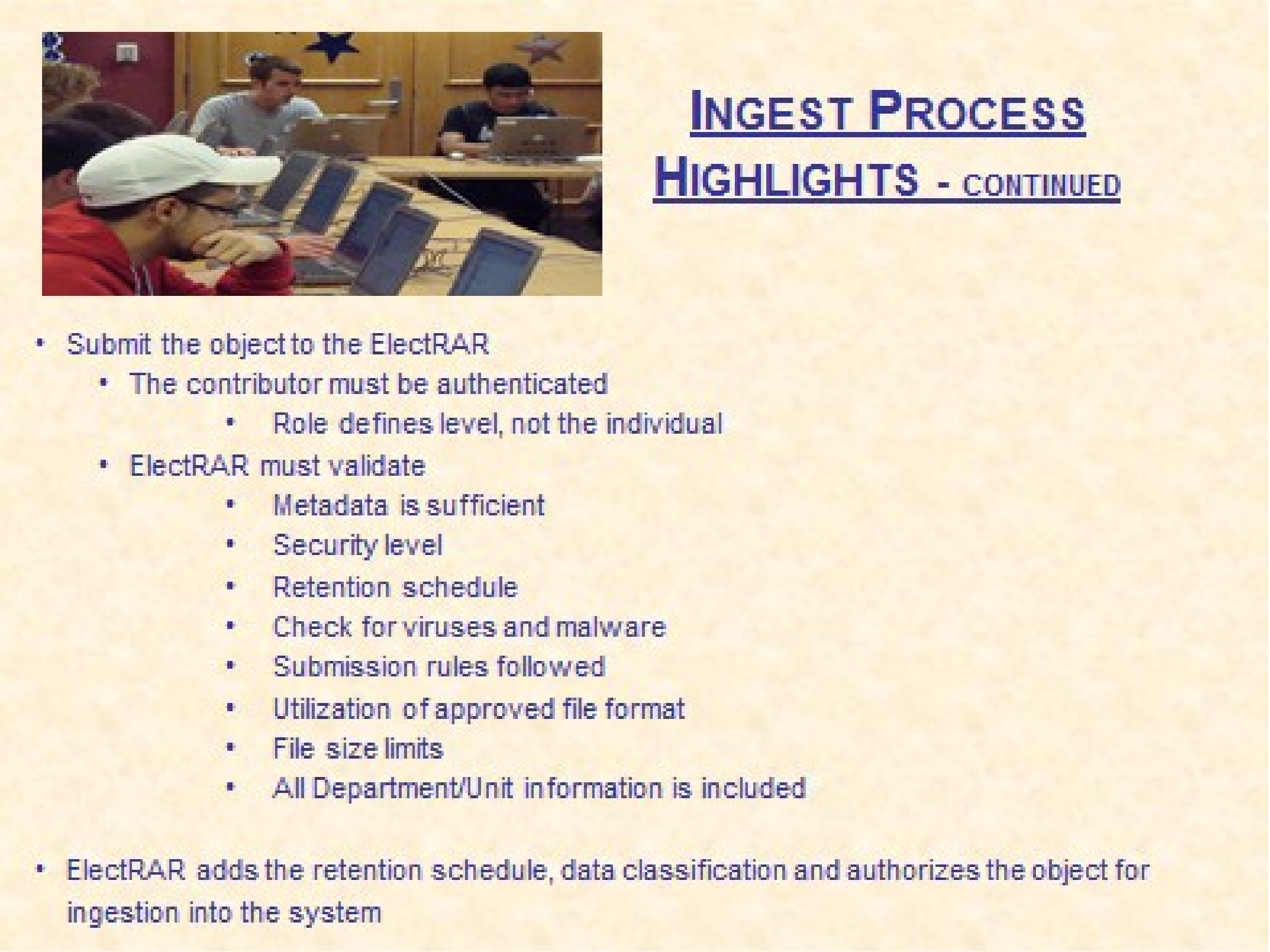
5.3 Authorize Information Submission

5.4
- Retention Schedules
- Data Classification

6 Proceed to E-Records Management Process

To Management Process

Ingest Process Highlights

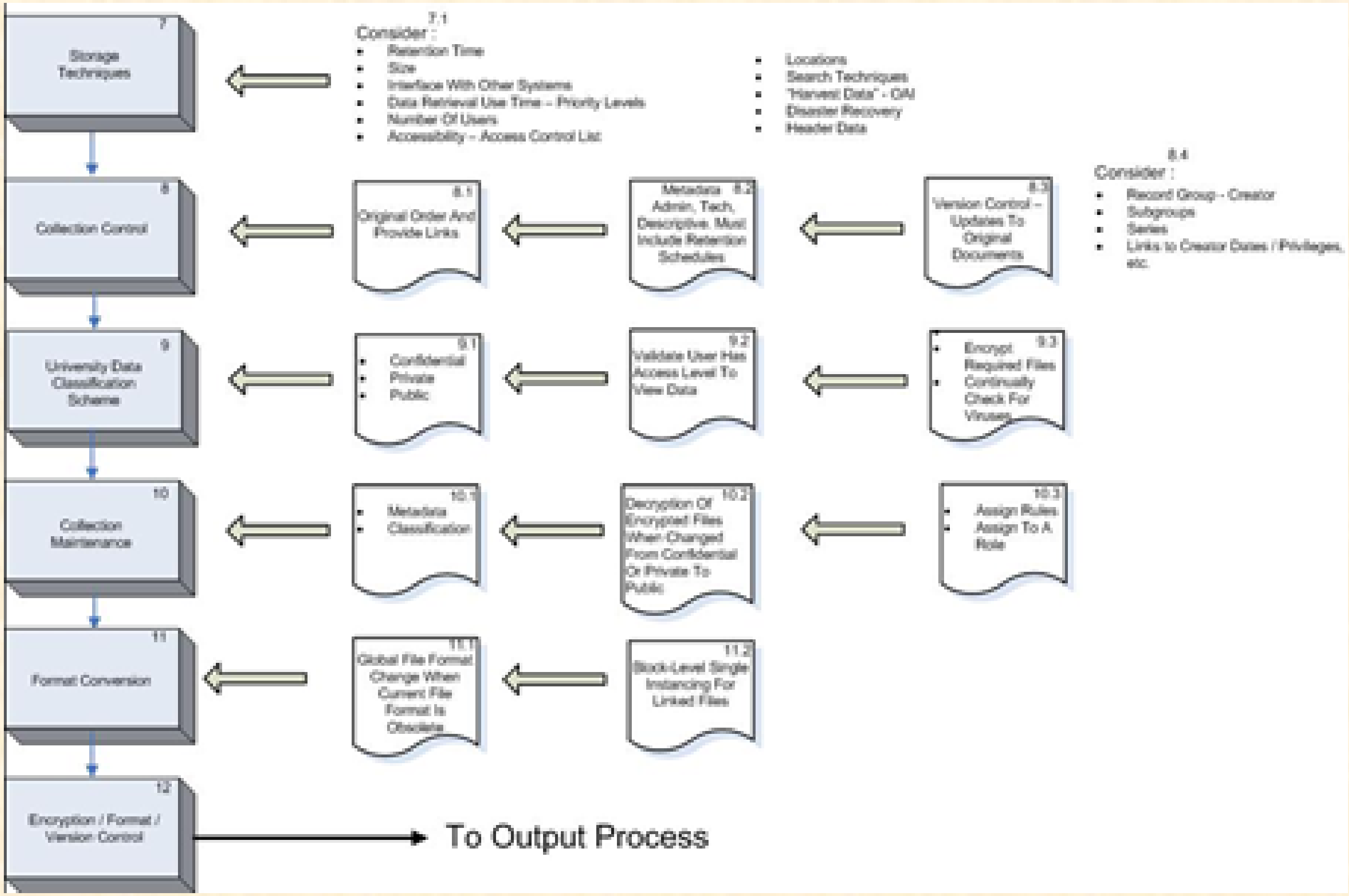


Ingest Process Highlights - continued

- Submit the object to the ElectRAR
  - The contributor must be authenticated
    - Role defines level, not the individual
  - ElectRAR must validate
    - Metadata is sufficient
    - Security level
    - Retention schedule
    - Check for viruses and malware
    - Submission rules followed
    - Utilization of approved file format
    - File size limits
    - All Department/Unit information is included
- ElectRAR adds the retention schedule, data classification and authorizes the object for ingestion into the system

Information Management Process

INFORMATION MANAGEMENT PROCESS



7 Storage Techniques

7.1 Consider :
- Retention Time
- Size
- Interface With Other Systems
- Data Retrieval Use Time – Priority Levels
- Number Of Users
- Accessibility – Access Control List
- Locations
- Search Techniques
- "Harvest Data" - OAI
- Disaster Recovery
- Header Data

8 Collection Control

8.1 Original Order And Provide Links

8.2 Metadata Admin, Tech, Descriptive, Must Include Retention Schedules

8.3 Version Control – Updates To Original Documents

8.4 Consider :
- Record Group - Creator
- Subgroups
- Series
- Links to Creator Dates / Privileges, etc.

9 University Data Classification Scheme

9.1
- Confidential
- Private
- Public

9.2 Validate User Has Access Level To View Data

9.3
- Encrypt Required Files
- Continually Check For Viruses

10 Collection Maintenance

10.1
- Metadata
- Classification

10.2 Decryption Of Encrypted Files When Changed From Confidential Or Private To Public

10.3
- Assign Rules
- Assign To A Role

11 Format Conversion

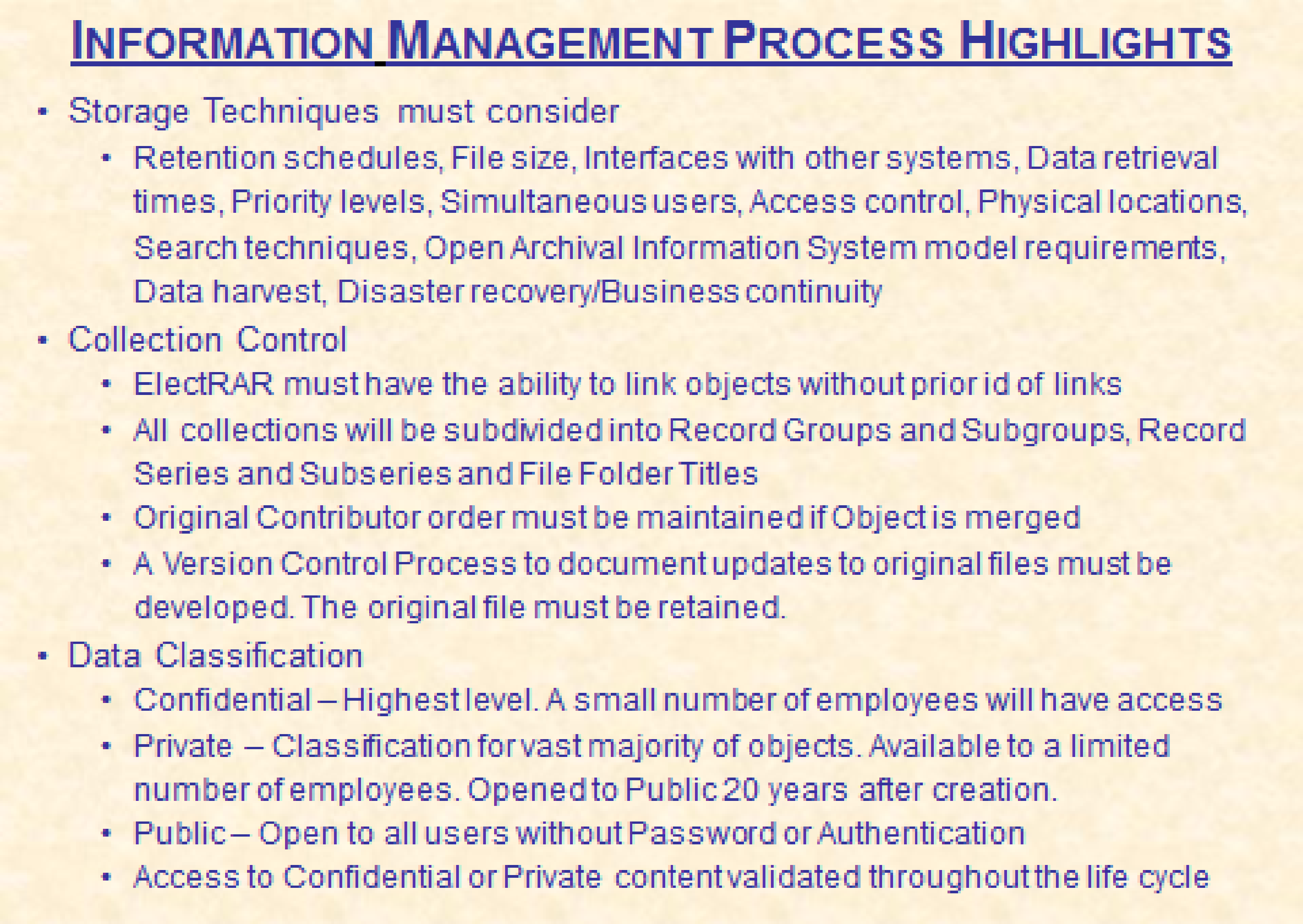
11.1 Global File Format Change When Current File Format Is Obsolete

11.2 Block-Level Single Instancing For Linked Files

12 Encryption / Format / Version Control

To Output Process

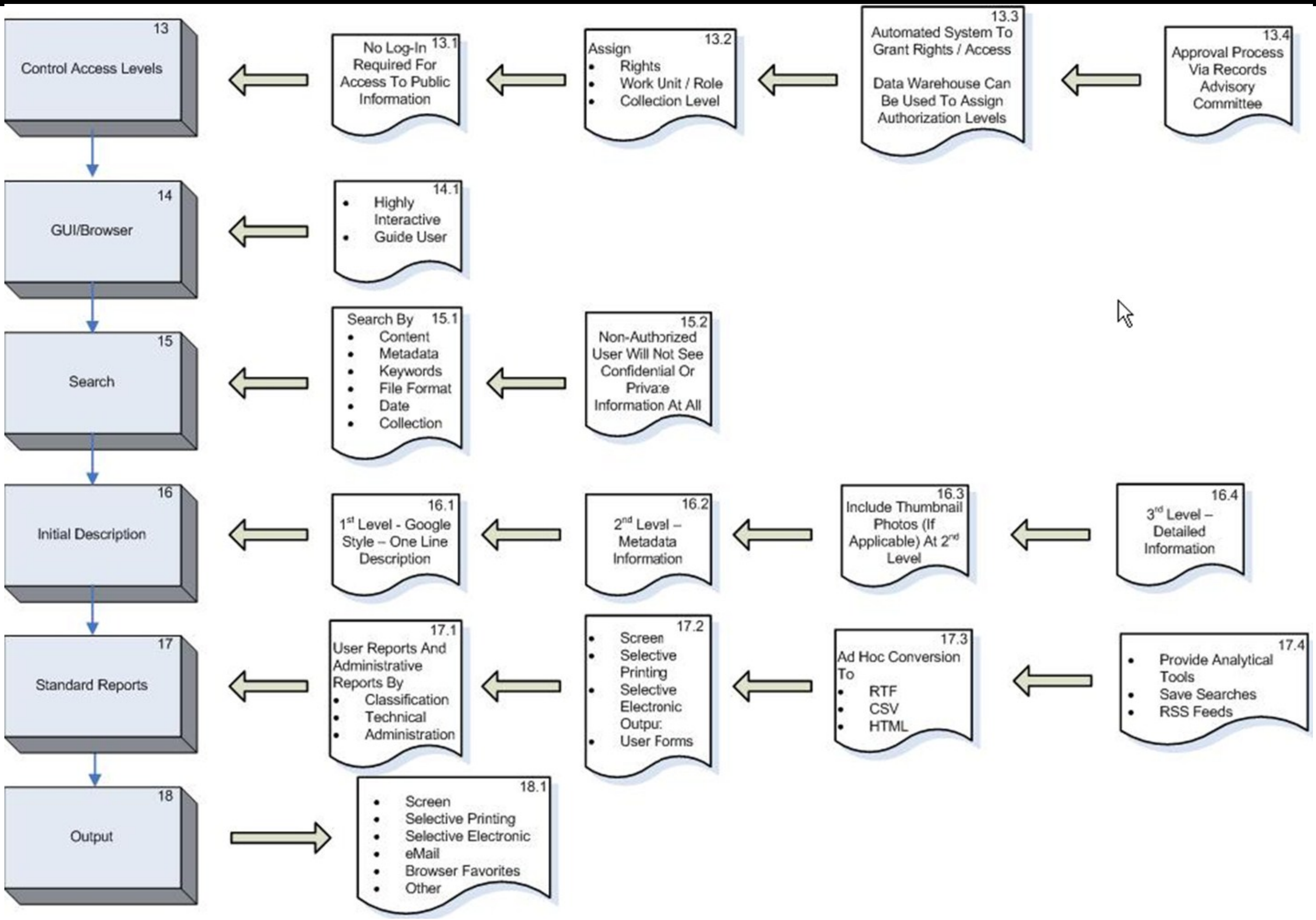
Information Management Process Highlights



INFORMATION MANAGEMENT PROCESS HIGHLIGHTS

- Storage Techniques must consider
  - Retention schedules, File size, Interfaces with other systems, Data retrieval times, Priority levels, Simultaneous users, Access control, Physical locations, Search techniques, Open Archival Information System model requirements, Data harvest, Disaster recovery/Business continuity
- Collection Control
  - ElectRAR must have the ability to link objects without prior id of links
  - All collections will be subdivided into Record Groups and Subgroups, Record Series and Subseries and File Folder Titles
  - Original Contributor order must be maintained if Object is merged
  - A Version Control Process to document updates to original files must be developed. The original file must be retained.
- Data Classification
  - Confidential – Highest level. A small number of employees will have access
  - Private – Classification for vast majority of objects. Available to a limited number of employees. Opened to Public 20 years after creation.
  - Public – Open to all users without Password or Authentication
  - Access to Confidential or Private content validated throughout the life cycle

Output Process Workflow



13 Control Access Levels

13.1 No Log-In Required For Access To Public Information

13.2 Assign
- Rights
- Work Unit / Role
- Collection Level

13.3 Automated System To Grant Rights / Access

Data Warehouse Can Be Used To Assign Authorization Levels

13.4 Approval Process Via Records Advisory Committee

14 GUI/Browser

14.1
- Highly Interactive
- Guide User

15 Search

15.1 Search By
- Content
- Metadata
- Keywords
- File Format
- Date
- Collection

15.2 Non-Authorized User Will Not See Confidential Or Private Information At All

16 Initial Description

16.1 1st Level - Google Style – One Line Description

16.2 2nd Level – Metadata Information

16.3 Include Thumbnail Photos (If Applicable) At 2nd Level

16.4 3rd Level – Detailed Information

17 Standard Reports

17.1 User Reports And Administrative Reports By
- Classification
- Technical
- Administration

17.2
- Screen
- Selective Printing
- Selective Electronic Output
- User Forms

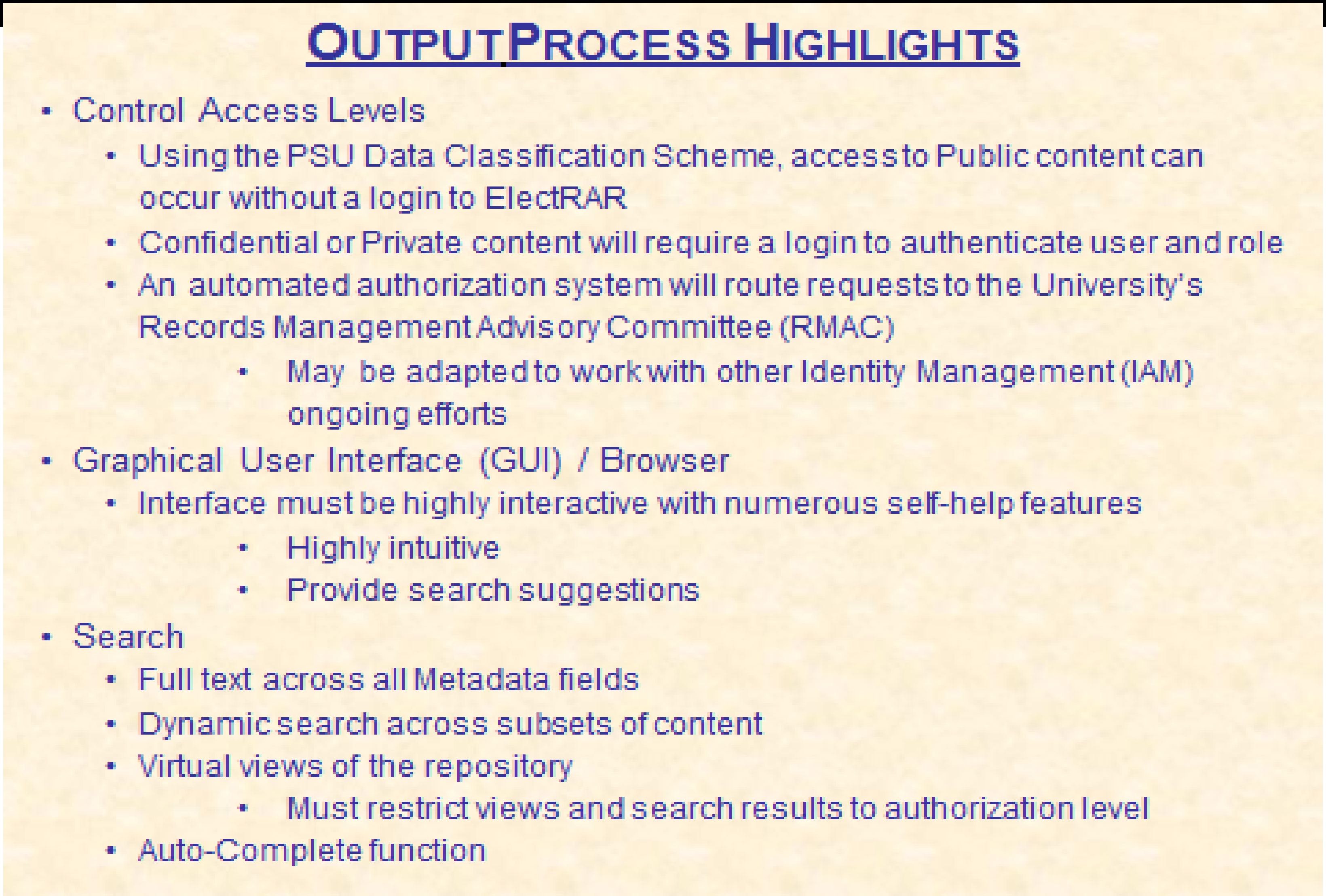
17.3 Ad Hoc Conversion To
- RTF
- CSV
- HTML

17.4
- Provide Analytical Tools
- Save Searches
- RSS Feeds

18 Output

18.1
- Screen
- Selective Printing
- Selective Electronic
- eMail
- Browser Favorites
- Other

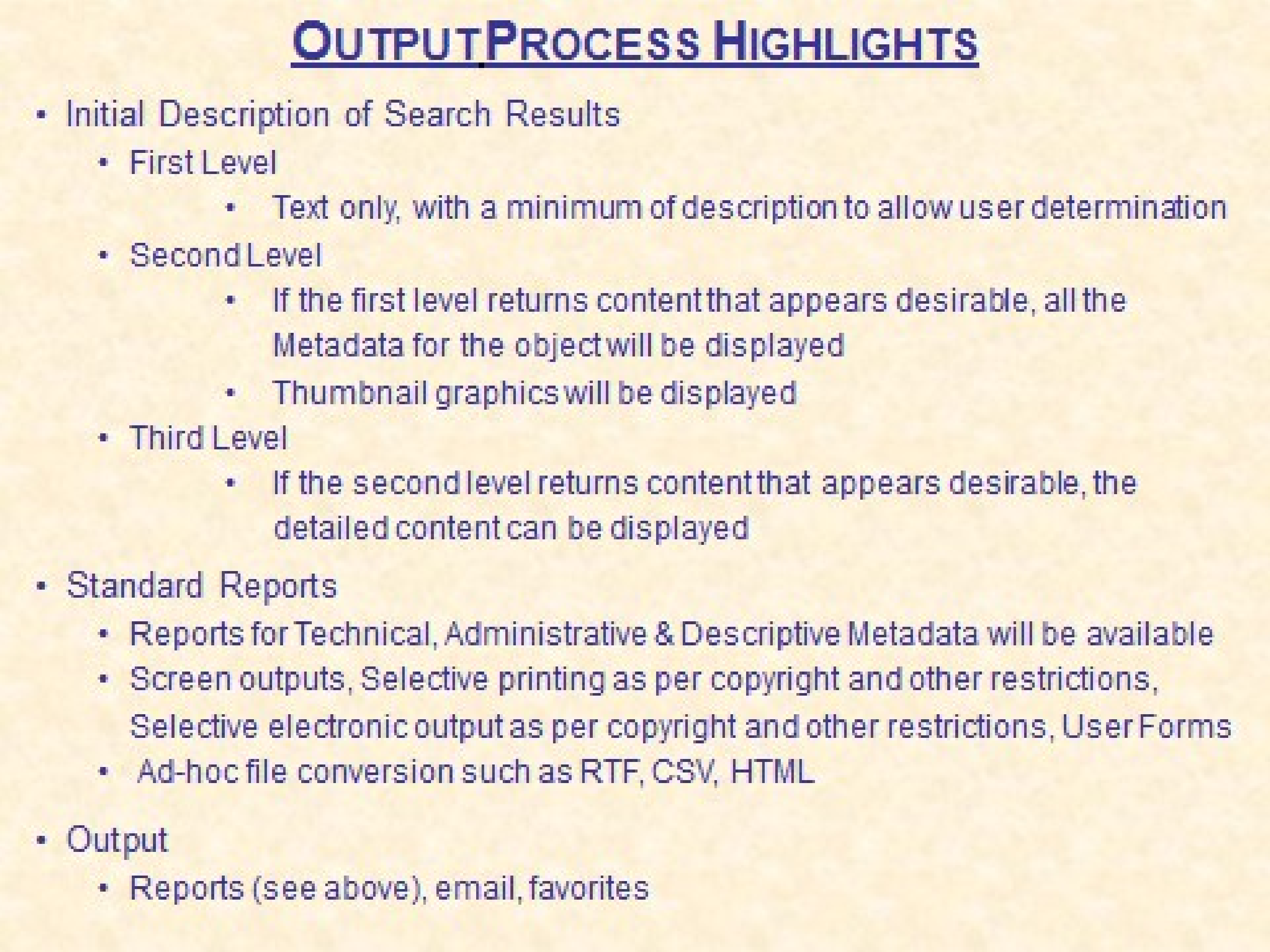
Output Process Highlights



OUTPUT PROCESS HIGHLIGHTS

- Control Access Levels
  - Using the PSU Data Classification Scheme, access to Public content can occur without a login to ElectRAR
  - Confidential or Private content will require a login to authenticate user and role
  - An automated authorization system will route requests to the University's Records Management Advisory Committee (RMAC)
    - May be adapted to work with other Identity Management (IAM) ongoing efforts
- Graphical User Interface (GUI) / Browser
  - Interface must be highly interactive with numerous self-help features
    - Highly intuitive
    - Provide search suggestions
- Search
  - Full text across all Metadata fields
  - Dynamic search across subsets of content
  - Virtual views of the repository
    - Must restrict views and search results to authorization level
  - Auto-Complete function

Output Process Highlights



OUTPUT PROCESS HIGHLIGHTS

- Initial Description of Search Results
  - First Level
    - Text only, with a minimum of description to allow user determination
  - Second Level
    - If the first level returns content that appears desirable, all the Metadata for the object will be displayed
    - Thumbnail graphics will be displayed
  - Third Level
    - If the second level returns content that appears desirable, the detailed content can be displayed
- Standard Reports
  - Reports for Technical, Administrative & Descriptive Metadata will be available
  - Screen outputs, Selective printing as per copyright and other restrictions, Selective electronic output as per copyright and other restrictions, User Forms
  - Ad-hoc file conversion such as RTF, CSV, HTML
- Output
  - Reports (see above), email, favorites

Then along came CAPS

- CURATION
- ARCHITECTURE
- PROTOTYPE
- SERVICES

CAPS - How it began

- 2009/10 Platform review

- -Four legacy systems:
  - CONTENTdm (mainly images, some text)
  - DPubS for journals & monographs
  - Olive ActivePaper Archive for historical newspapers
  - ETD database system for theses/dissertations

- No platform for electronic records (or research data)

Inefficiencies revealed

- Silos - different workflows, training & back end technology
- Focus on content delivery rather than management - centralized preservation impossible
- Information dispersed - some in applications, others in file systems, others in personal spreadsheets
- 3/4 delivery applications moribund and we don't have access to the source code

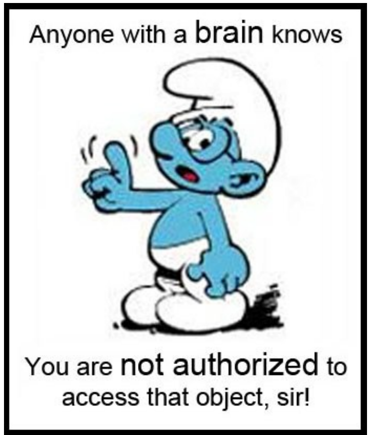
Curation Microservices Proof-of-Concept

- Team: 4 members from 3 departments
  - Digital Collections Curator
  - Digital Architect
  - Programmer
  - Archivist (moi)
- Digital Libraries Technology sent us all to a microservices (un)conference

What are micro-services?

- “Small things…specialized jobs…only truly powerful when they work in concert…ZOMG IT'S THE SMURFS” –Michael B. Klein
- Small, self-contained, independent services
- Easier to develop, deploy, maintain, enhance, replace.
- Interoperable: combine for more complex applications.

https://confluence.ucop.edu/display/Curation/Home


Anyone with a brain knows
You are not authorized to access that object, sir!

The micro-services "philosophy"

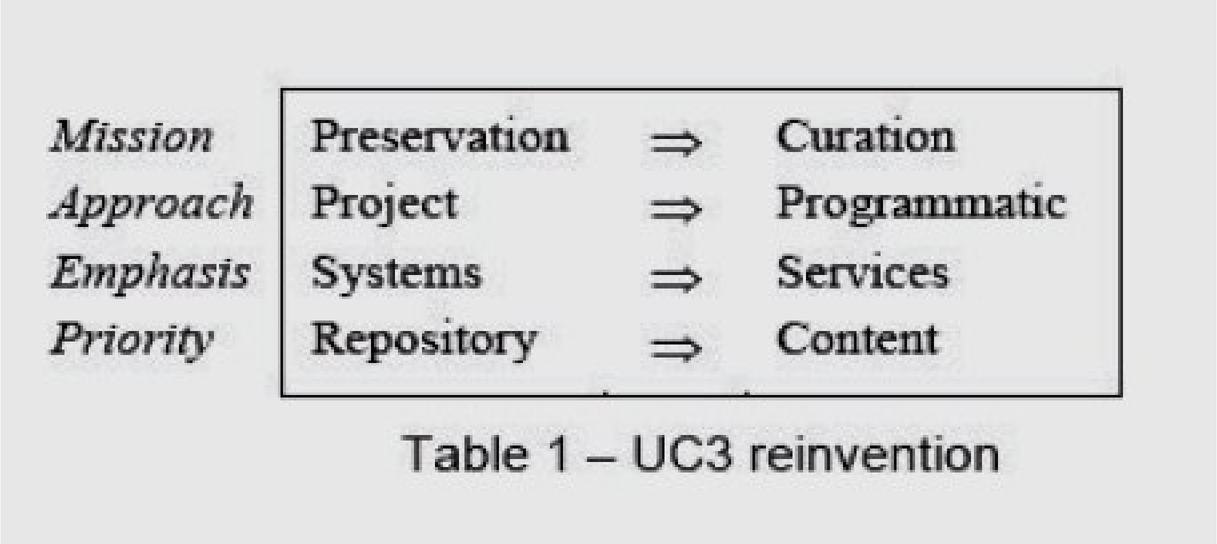
Metaphors	*Assumptions*	*Principles*	*Preferences*	*Practices*
Pipeline	Safety through redundancy	Modularity	The small and simple over the large and complex	Focus on outcomes, not means
Lego bricks	Meaning through context	Granularity	The minimally sufficient over the feature laden	Complexity through composition, not addition
	Utility through service	Orthogonality	The configurable over the prescribed	Policy neutral, platform and protocol independent
	Value through use (and reuse)	Emergence	The proven over the (merely) novel	Approach sufficiency through incrementally necessary steps
	Stewardship is a relay	Evolution		Early prototyping, frequent refactoring
		Parsimony		Code to interfaces

http://groups.google.com/group/digital-curation

Examples of microservices

- Annotate - describe or catalog an object
- Authenticate - authenticate a user
- Authorize - authorize a user to access an object
- Characterize - generate administrative metadata for an object
- Identify - generate an identifier for an object
- Inventory - record an object's location on disk
- Relate - relate two or more objects
- Store - store an object on a filesystem
- Verify - check the integrity/checksum/fixity of an object
- Version - add a version to an object

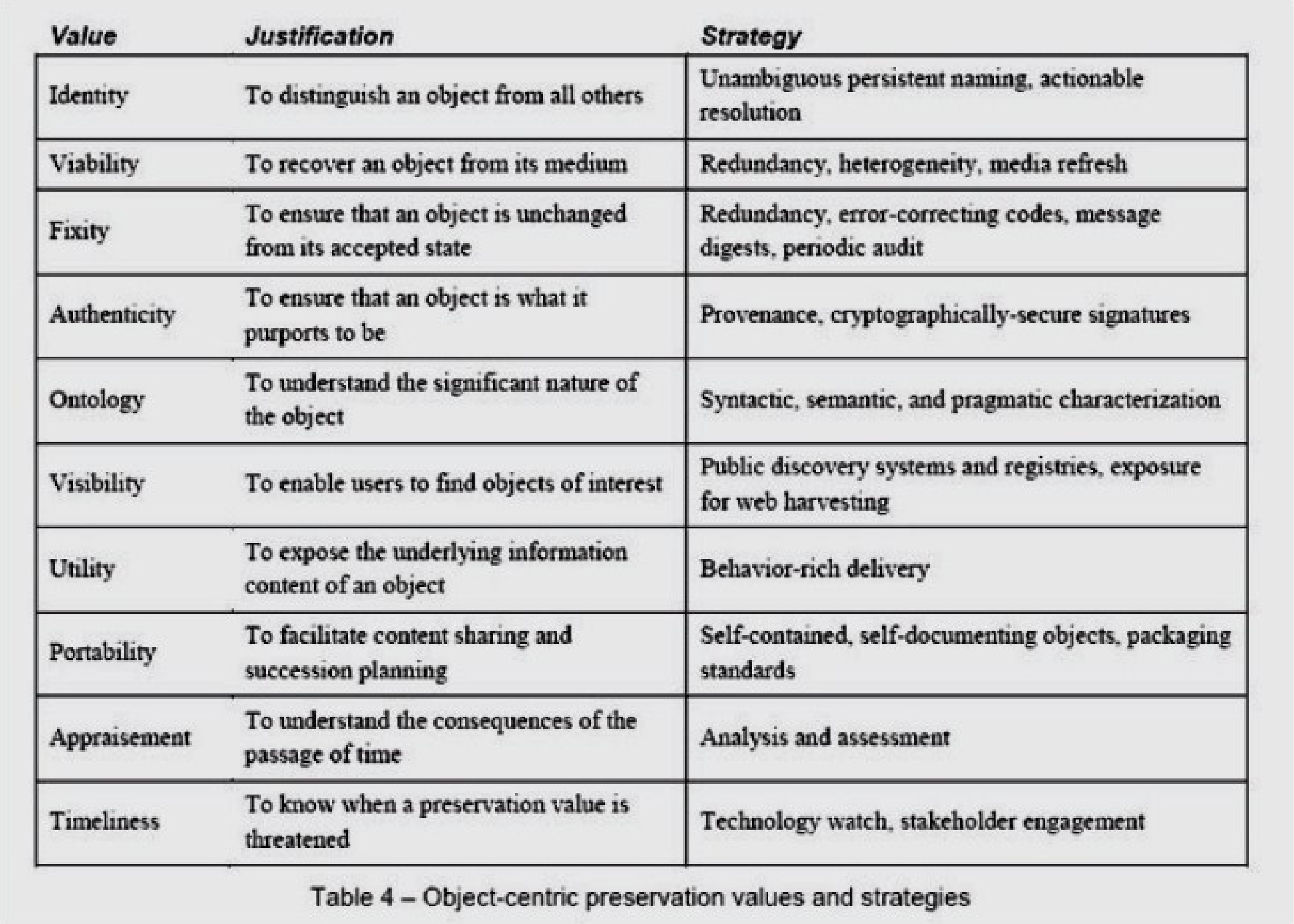
Credit to California Digital Library



Mission	Preservation	⇒	Curation
Approach	Project	⇒	Programmatic
Emphasis	Systems	⇒	Services
Priority	Repository	⇒	Content

Table 1 – UC3 reinvention

Based on curation values



Value	Justification	Strategy
Identity	To distinguish an object from all others	Unambiguous persistent naming, actionable resolution
Viability	To recover an object from its medium	Redundancy, heterogeneity, media refresh
Fixity	To ensure that an object is unchanged from its accepted state	Redundancy, error-correcting codes, message digests, periodic audit
Authenticity	To ensure that an object is what it purports to be	Provenance, cryptographically-secure signatures
Ontology	To understand the significant nature of the object	Syntactic, semantic, and pragmatic characterization
Visibility	To enable users to find objects of interest	Public discovery systems and registries, exposure for web harvesting
Utility	To expose the underlying information content of an object	Behavior-rich delivery
Portability	To facilitate content sharing and succession planning	Self-contained, self-documenting objects, packaging standards
Appraisement	To understand the consequences of the passage of time	Analysis and assessment
Timeliness	To know when a preservation value is threatened	Technology watch, stakeholder engagement

Table 4 – Object-centric preservation values and strategies

Penn State curators

- Before we could evaluate the microservices approach, we needed to know how *our* curators work

- So we gathered use cases from:
  - University Archives
  - Digitization & Preservation
  - Art and Architecture Library
  - Maps Library

Example Use Case

The head of an academic department is complaining to the Provost that he did not approve a course currently being taught by a new professor in his department.

Course proposals must pass through 3 levels of approval. Course proposals are archived in digital format, and the three layers of approval are recorded through digital signatures.

The Provost asks the University Archivist to retrieve the course proposal and verify that the department head signed off on it. The course proposal shows that indeed it went through all appropriate approvals. The University Archivist must make the case that the department head's (digital) signature is authentic.

The University Archivist must also make sure that the version of the course proposal signed off on by the department head is the same version currently being taught.

From Proof-of-Concept to Prototype

Our explorations were positive enough that we were asked to take the next step

Added Asst. Head of IT (as project manager) & Metadata Librarian (from cataloging) to team

With "stakeholders" from 4 additional departments/ libraries, now 9 units represented)

The **relate** tool

Worked better than expected

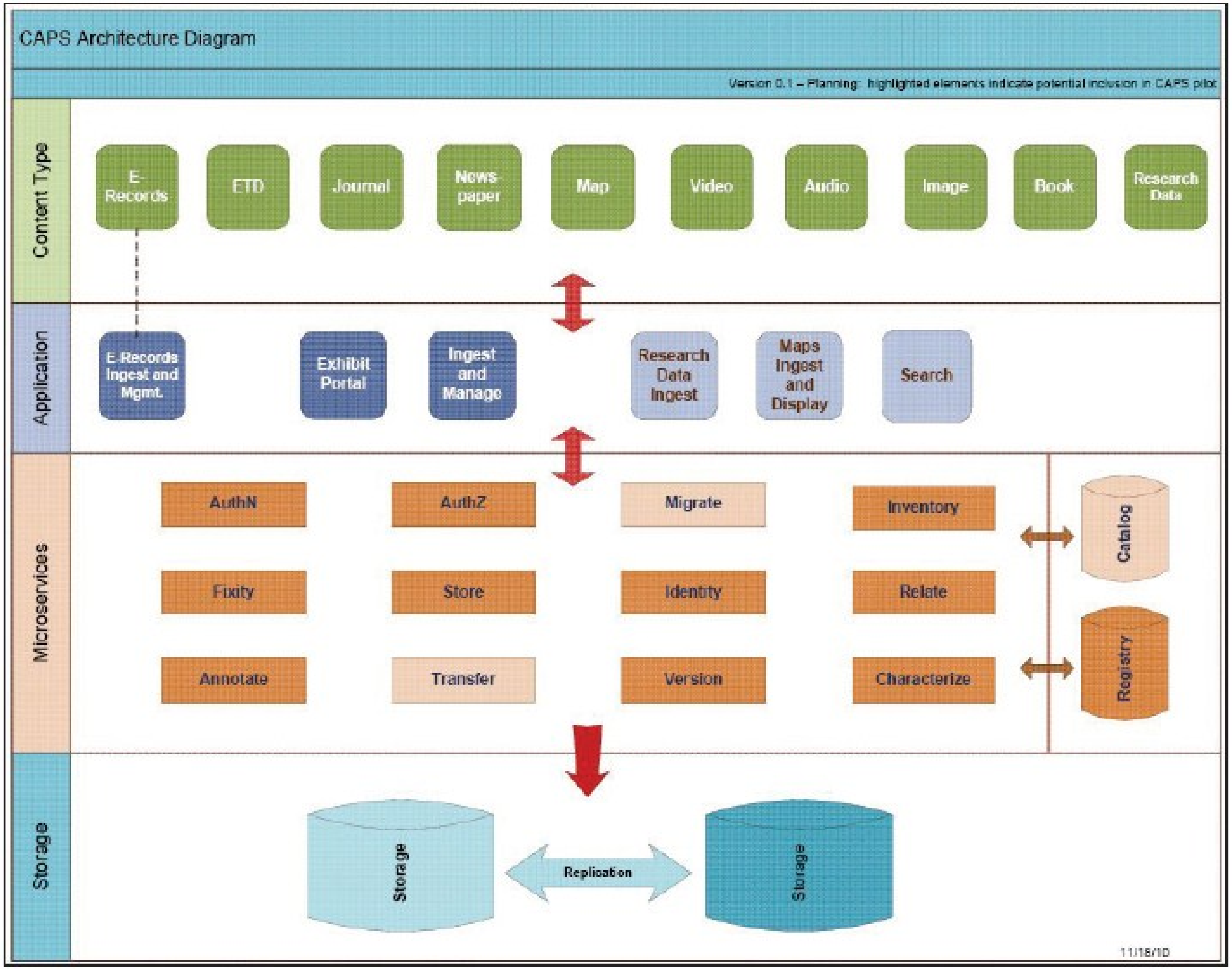
Goals of CAPS team

- Develop, test & assess a curation services architecture
- Engage library staff in development of applications for ingest, management and retrieval/delivery
- Apply agile development practices
- Document experiences and share source code

Process (outreach & agility)

- Daily meetings with core team
- Weekly meetings with stakeholders
- Constantly incorporating feedback into our work and reformulating long/short term goals
- Never “no” – just “not now”
- Progress tracked immediately on wiki
- Led to buy-in from stakeholders as well as improved final product
- Developed prototype product in 3 months time

CAPS architecture



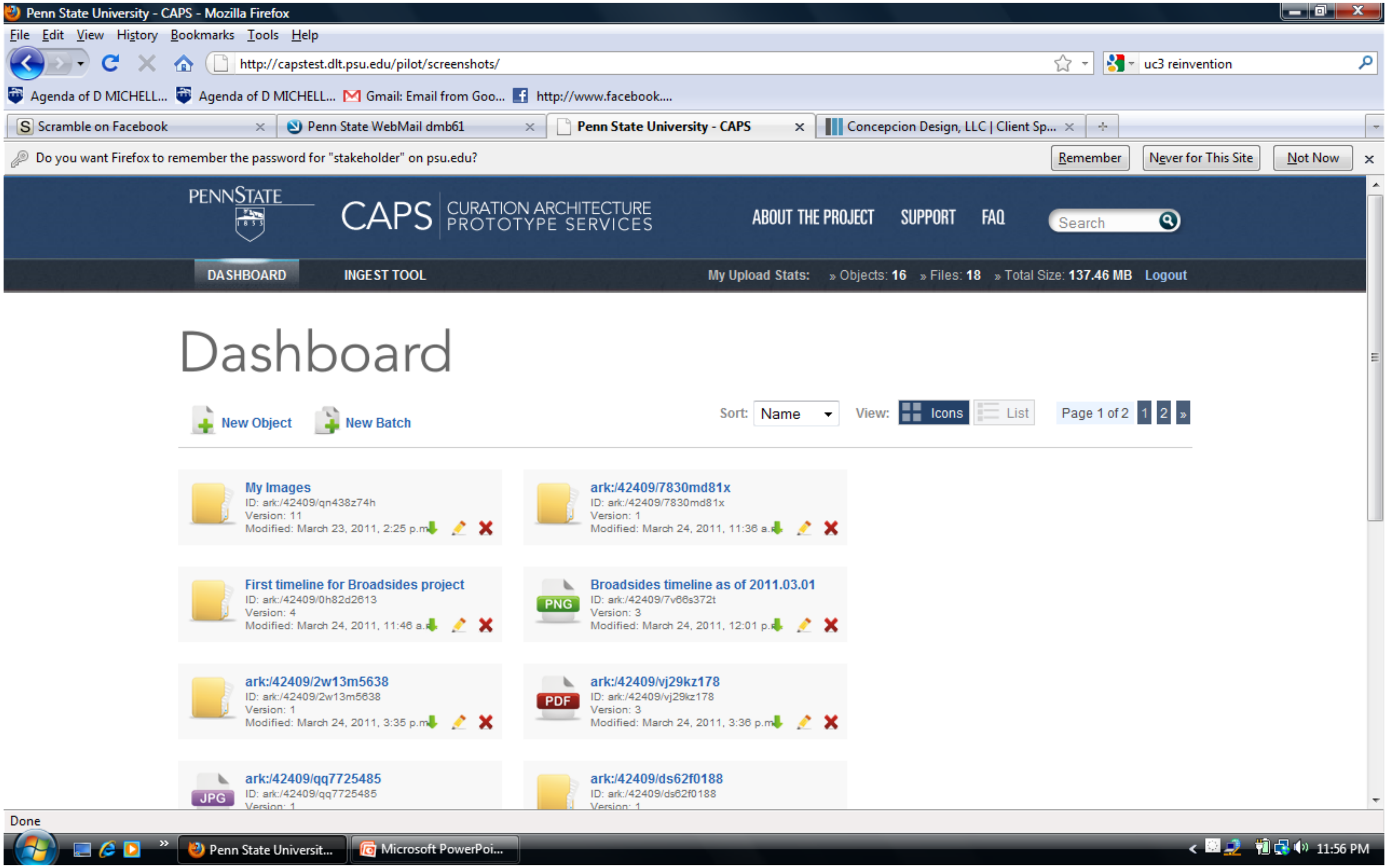
Open Source

- All of the software libraries and tools that power CAPS are released as open source, such as Python, Git, Django, jQuery, and MySQL.
- Benefits:
  - Aligning our development efforts with the broader technology community.
  - Build on existing code
  - Rapid identification and resolution of bugs
  - Experience collaborating with peer institutions

Metadata

- Phase One objectives:
  - Survey stakeholders for their needs
  - Derive a simple, extensible standard to underlie the system's search functions.
- Currently using Dublin Core, modeled using the Resource Description Framework (RDF), allowing for interoperability
- Data dictionary to outline the fields currently in the system
- Will grow to include necessary preservation, technical, and administrative metadata fields, as the processes for collecting them become more specific in future phases

Screenshot (1)



Penn State University - CAPS - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Agenda of D MICHELL... Agenda of D MICHELL... Gmail: Email from Goo... http://www.facebook....
Scramble on Facebook
Penn State WebMail dmb61
Penn State University - CAPS
Concepcion Design, LLC | Client Sp...
Remember
Never for This Site
Not Now
PENNSTATE
CAPS CURATION ARCHITECTURE PROTOTYPE SERVICES
ABOUT THE PROJECT
SUPPORT
FAQ
Search
DASHBOARD
INGEST TOOL
Logout
New Object
New Batch
Icons
List
1 2 »
Done
Penn State Universit...
Microsoft PowerPoi...
11:56 PM

http://capstest.dlt.psu.edu/pilot/screenshots/

uc3 reinvention

Do you want Firefox to remember the password for "stakeholder" on psu.edu?

My Upload Stats: » Objects: 16 » Files: 18 » Total Size: 137.46 MB

Dashboard

Sort: Name

View: Icons

Page 1 of 2

My Images
ID: ark:/42409/qn438z74h
Version: 11
Modified: March 23, 2011, 2:25 p.m

ark:/42409/7830md81x
ID: ark:/42409/7830md81x
Version: 1
Modified: March 24, 2011, 11:36 a.

First timeline for Broadsides project
ID: ark:/42409/0h82d2613
Version: 4
Modified: March 24, 2011, 11:46 a.

Broadsides timeline as of 2011.03.01
PNG
ID: ark:/42409/7v66s372t
Version: 3
Modified: March 24, 2011, 12:01 p.

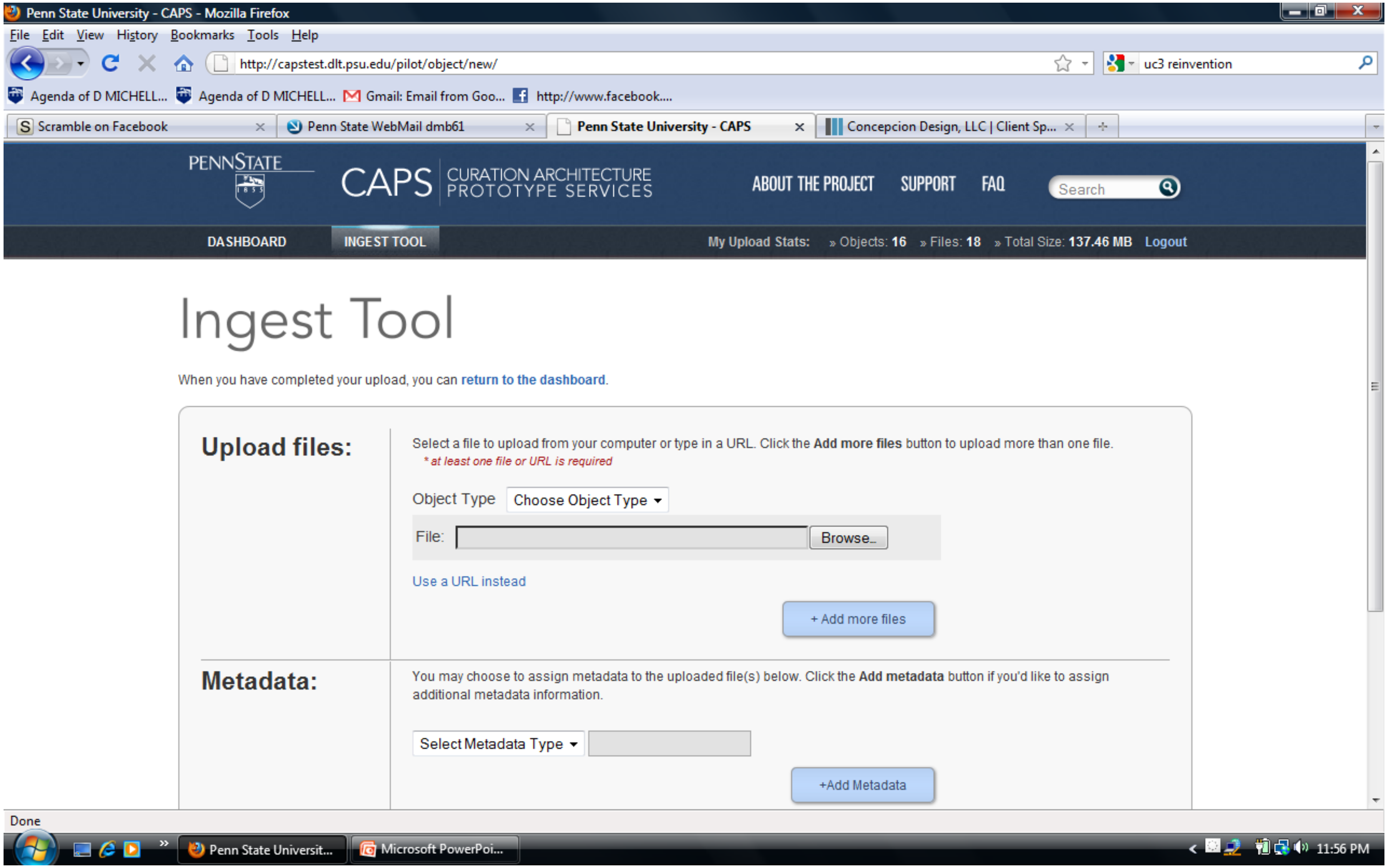
ark:/42409/2w13m5638
ID: ark:/42409/2w13m5638
Version: 1
Modified: March 24, 2011, 3:35 p.m

ark:/42409/vj29kz178
PDF
ID: ark:/42409/vj29kz178
Version: 3
Modified: March 24, 2011, 3:36 p.m

ark:/42409/qq7725485
JPG
ID: ark:/42409/qq7725485
Version: 1

ark:/42409/ds62f0188
ID: ark:/42409/ds62f0188
Version: 1

Screenshot (2)



Penn State University - CAPS - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://capstest.dlt.psu.edu/pilot/object/new/
uc3 reinvention
Agenda of D MICHELL... Agenda of D MICHELL... Gmail: Email from Goo... http://www.facebook....
Scramble on Facebook
Penn State WebMail dmb61
Penn State University - CAPS
Concepcion Design, LLC | Client Sp...
PENNSTATE
CAPS CURATION ARCHITECTURE PROTOTYPE SERVICES
ABOUT THE PROJECT
SUPPORT
FAQ
Search
DASHBOARD
INGEST TOOL
Logout
Done
Penn State Universit...
Microsoft PowerPoi...
11:56 PM

My Upload Stats: » Objects: **16** » Files: **18** » Total Size: **137.46 MB**

Ingest Tool

When you have completed your upload, you can **return to the dashboard**.

Upload files:

Select a file to upload from your computer or type in a URL. Click the **Add more files** button to upload more than one file.

** at least one file or URL is required*

Object Type: Choose Object Type

File: [Browse...]

Use a URL instead

[+ Add more files]

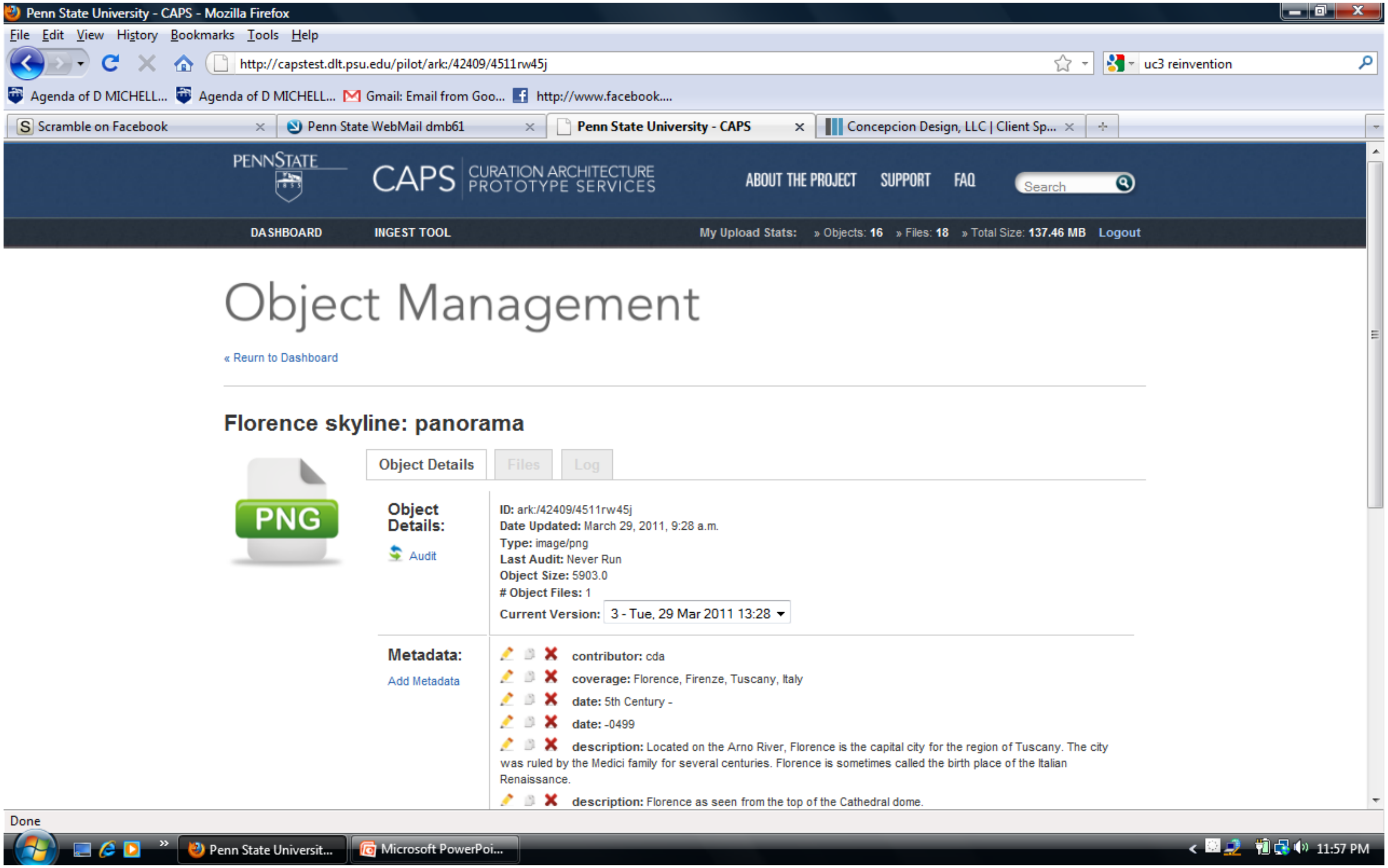
Metadata:

You may choose to assign metadata to the uploaded file(s) below. Click the **Add metadata** button if you'd like to assign additional metadata information.

Select Metadata Type

[+Add Metadata]

Screenshot (3)



My Upload Stats: » Objects: 16 » Files: 18 » Total Size: 137.46 MB

Object Management

« Reurn to Dashboard

Florence skyline: panorama

Object Details:

ID: ark:/42409/4511rw45j
Date Updated: March 29, 2011, 9:28 a.m.
Type: image/png
Last Audit: Never Run
Object Size: 5903.0
Object Files: 1
Current Version: 3 - Tue, 29 Mar 2011 13:28

Metadata:

contributor: cda
coverage: Florence, Firenze, Tuscany, Italy
date: 5th Century -
date: -0499
description: Located on the Arno River, Florence is the capital city for the region of Tuscany. The city was ruled by the Medici family for several centuries. Florence is sometimes called the birth place of the Italian Renaissance.
description: Florence as seen from the top of the Cathedral dome.

Assessment

- Survey of stakeholders in March - all agreed or strongly agreed that:
  - The team did a good job of listening to stakeholder concerns,
  - They were pleased with the prioritization of requirements, and
  - The mock-ups of the project deliverables reflected the prioritized stakeholder requirements
- One of the best outcomes is the building of community

  (9 different departments involved)

Next Steps

- Server Space -- $175,000 provided by the Vice Provost for Information Technology
- Start experimenting with existing tools and services for ingest, as architecture is developed
- That development will include:
  - E-records specific metadata
  - Retention periods
  - Levels of access
- Special Collections getting E-Records Archivist

Challenges

- Multiple sets of expectations
- Varied & shifting administrative priorities
- Staff time
- Once community is built, it has to be maintained – CAPS wrapped in March, still waiting on clear picture of development timeline

Anticipated Future --

- University Libraries Centralized Service called ***OpenCASA***: Curatorial Archival Services and Architecture; a component of which will be the
- ***Lion's Lair:* Libraries' Archival Information Repository**