Digital Preservation: Theory Approaching Practic

Richard Fyffe
Asst. Dean of Libraries for Scholarly Communication
Deborah Ludwig
Director, Enterprise Academic Systems
Beth Forrest Warner
Asst. Vice Provost for Information Services
University of Kansas

CNI Fall Meeting, Portland, OR *December 7, 2*

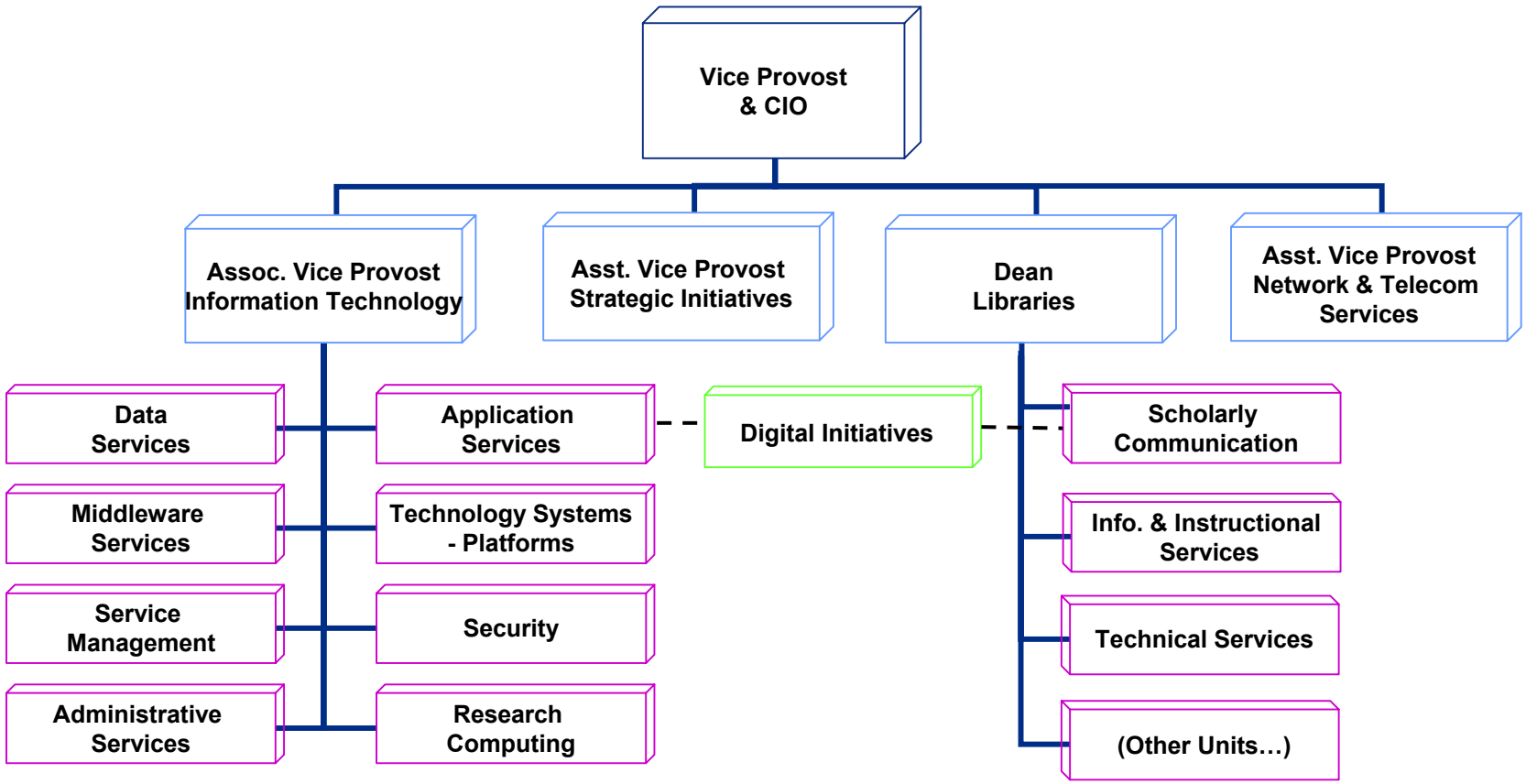
Presentation Overview

- Organizational Context
- Preservation Lifecycle Management
- Technical Architecture
- Roles and Policies
- Educational Outreach

KU Information Services Strategic Context

- HVC^2 Projects: *High Velocity Change through High Volume Collaboration*
  - *Collaborative Learning Spaces* - improving campus learning centers by better integrating print and electronic resources;
  - *Quality of Service Models* for students, scholars, and decision makers - designing new service models &delivery methods;
  - *Digital Preservation* - preserving digital resources in administrative, instruction, and scholarly research areas
- Institutional Repository: *KU ScholarWorks*
- Information Services Organizational Structure

Information Services: Organizational Structure



Digital Preservation Task Force Charge

Explore the implications of a University commitment to the preservation of digital assets:

- Define what “digital preservation” means in the context of University commitment and stewardship.
- Identify existing and potential University digital assets.
- Recommend criteria for selecting the assets for which the University will (and will not) take responsibility.
- Characterize a digital preservation infrastructure: organizational, technical, policy, financial.

Digital Preservation Task Force Membership

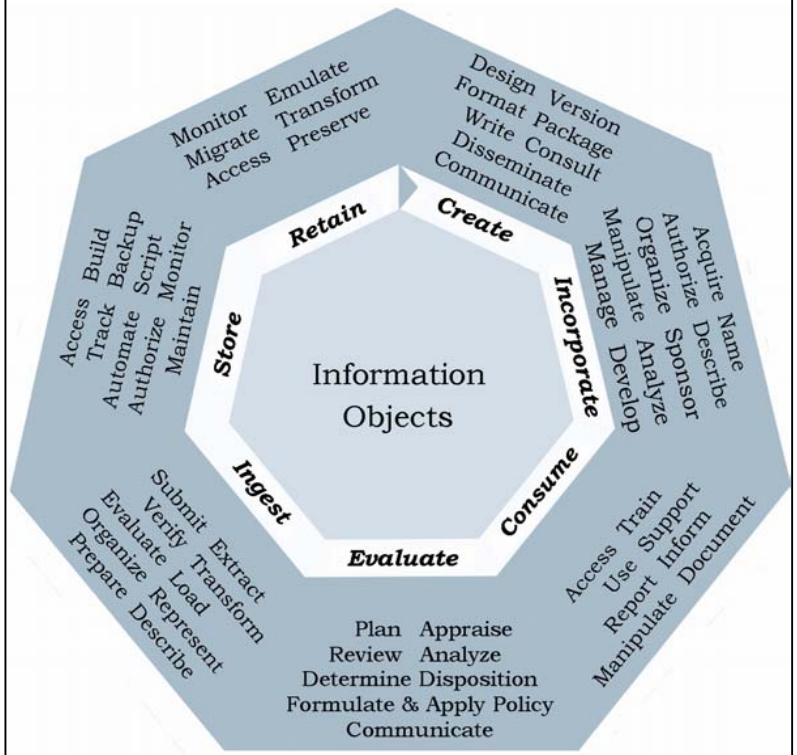
- Dean of Fine Arts (co-chair)
- Vice-Provost for Information Services
- Associate Vice-Provost for Information Services
- Assistant Vice-Provost for Information Services (Strategic Initiatives)
- Director of Academic Systems (Information Technology)
- Libraries:
  - Assistant Dean for Scholarly Communication (co-chair)
  - Assistant Dean for Technical Services
  - Preservation Officer
  - University Archivist
- Associate Director, Office of Institutional Research and Planning
- Assistant to the Associate Vice Chancellor for Information Resources, KU Medical Center
- Facilitators

Overview

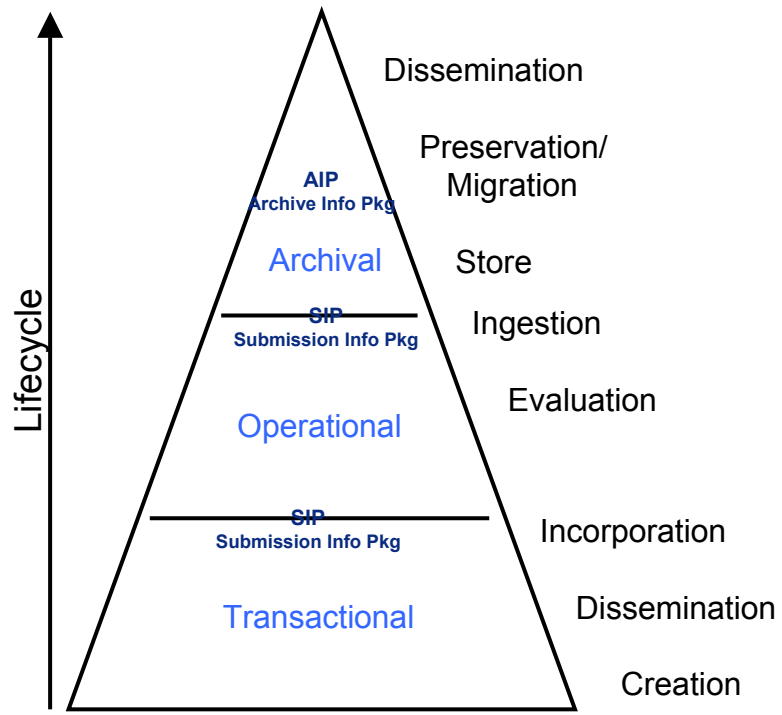
- Digital preservation is the **ongoing process of managing data for continuing access and use.** It is an outcome of the organization's successful day-to-day management of its digital information.
- ***Recommendation:***
  - **university-wide digital preservation program**
    - administrative & academic data
  - **three-year timeframe**
  - **implementation of the following components:**
    - Integrated **technical architecture** designed around the **whole lifecycle** of digital information
    - Definition and assignment of a set of specific **roles** or functions exercised by staff within the University, and development of a set of **policies** to guide those roles.
    - **Education** for faculty, staff, and administrators in the basic concepts and challenges in digital preservation and **training** in information management practices that will contribute to the ongoing availability of digital files.

Lifecycle Management for Digital Assets: Two Dimensions

Functional Dimension



Chronological Dimension



Digital Object Process / Lifecycle / Datastore Overview

Special Focus on the Upstream End of the Cycle

- Appraisal
- Choice of file format
- Choice of storage location / repository
- Metadata

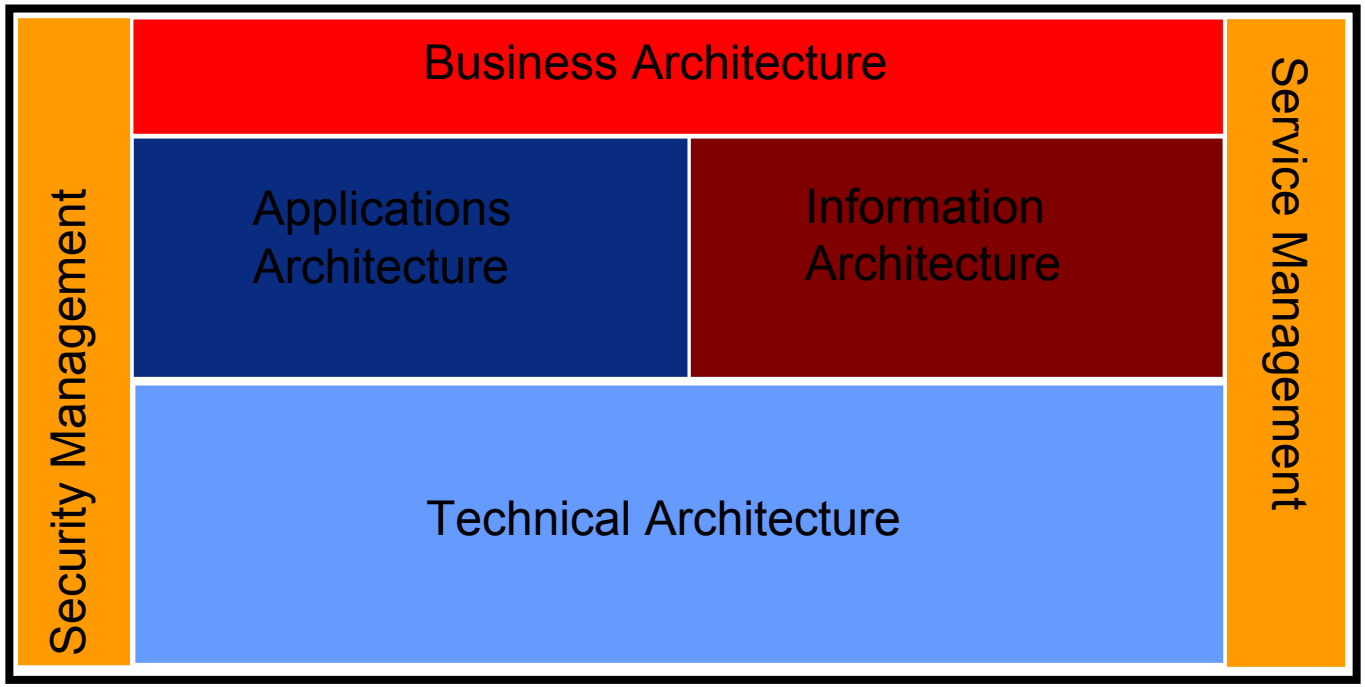
Consultation

Education

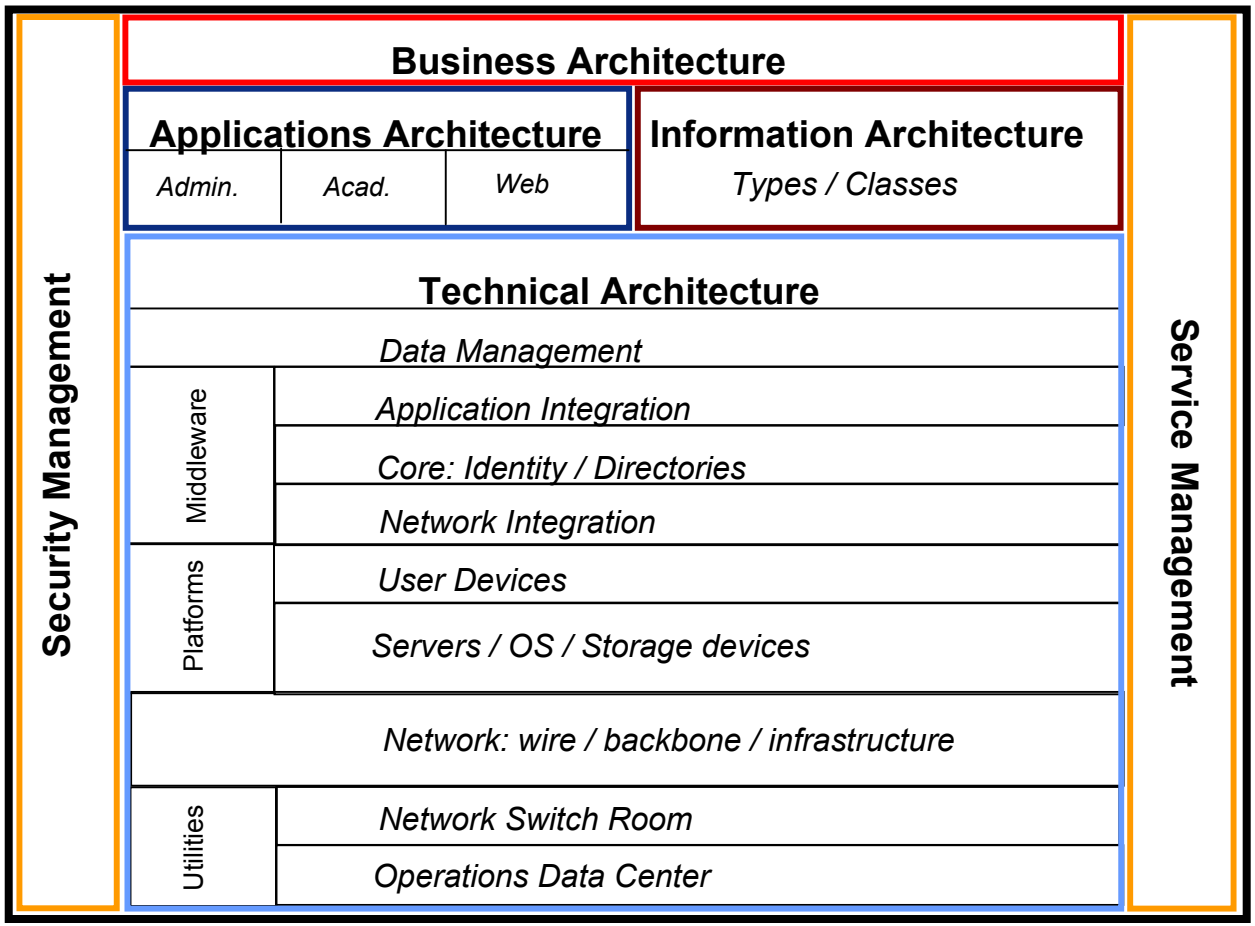
Infrastructure & Architecture

- Digital preservation infrastructure cannot be defined separately from the enterprise's basic technical infrastructure – it must be defined within a consistent technical architecture for the enterprise
- Architecture guides the overall development and establishes consistency by helping to:
  - provide a mechanism for a constant view of the information system infrastructure to serve as the basis from which the various groups of IT professionals develop and deliver information systems and services
  - provide business support services managers and staff with an understanding of the information systems infrastructure they are using; and
  - ensure that external development projects or application packages do not make incompatible changes to the infrastructure

KU Enterprise Architecture Framework



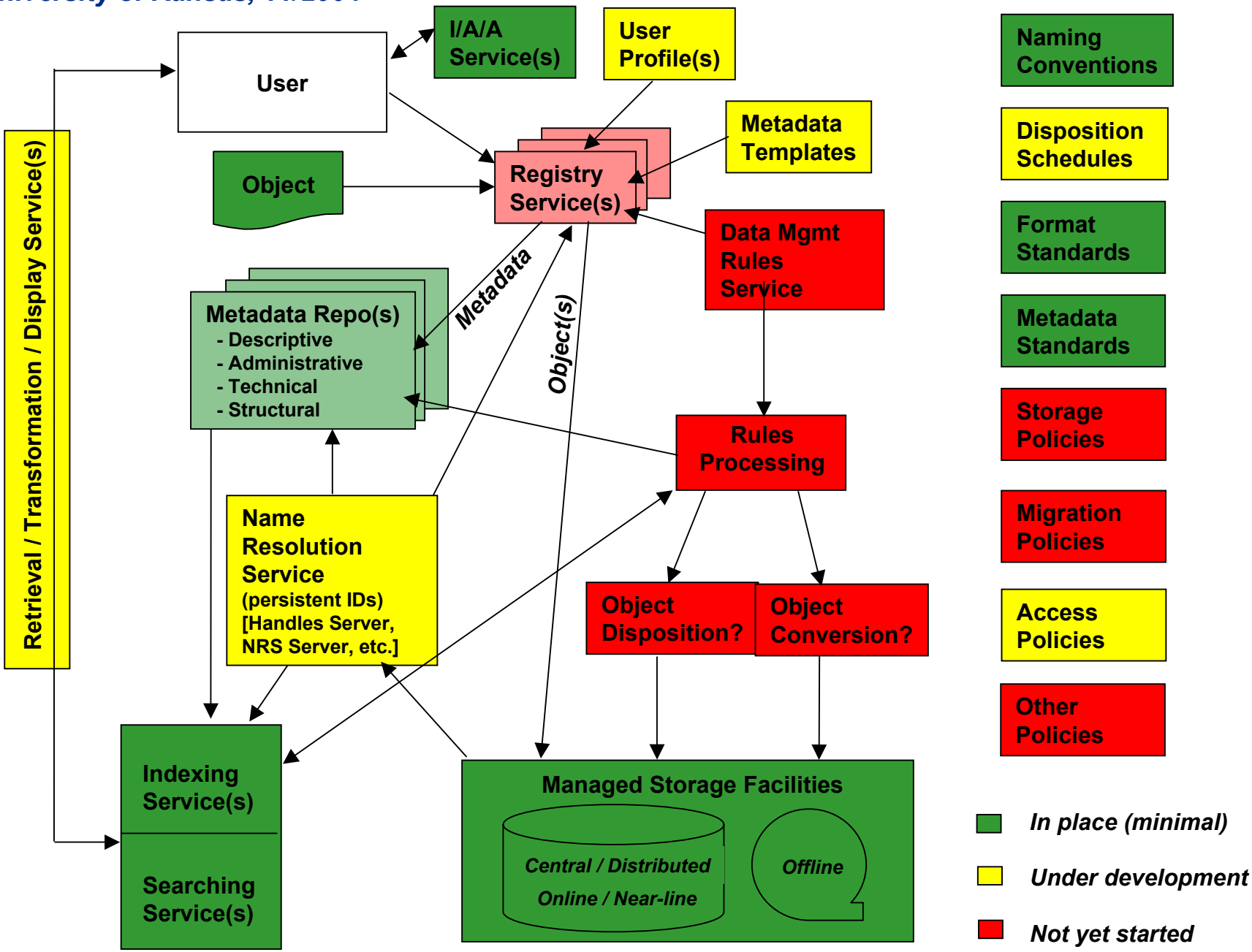
KU Information Services Organizational Framework



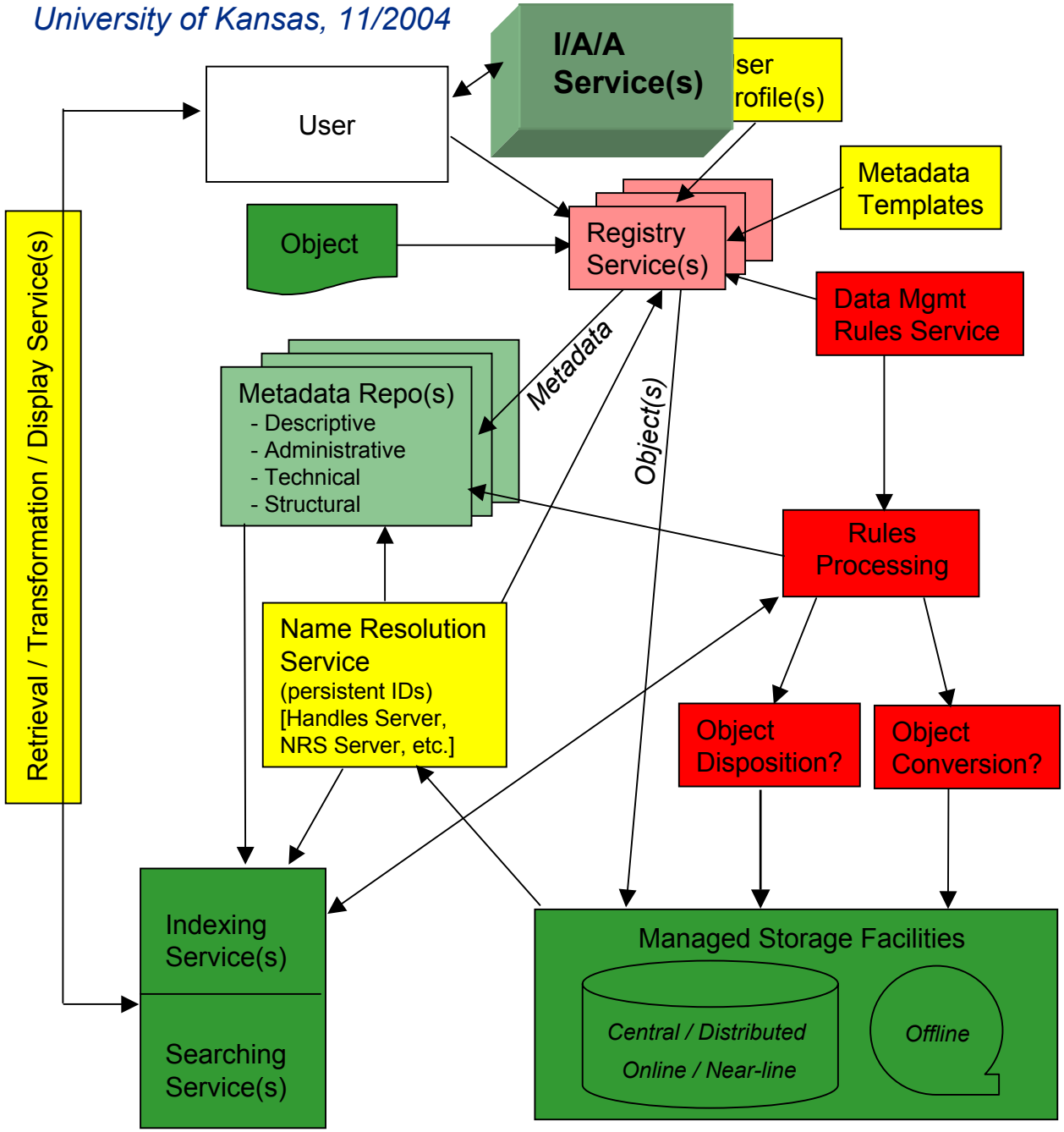
Technical Infrastructure

Within the architectural framework, infrastructure components that contribute to lifecycle information management can be identified…

University of Kansas, 11/2004



University of Kansas, 11/2004



Identification / Authentication / Authorization Services:

Services that:

- identify and verify users and their membership in a specific community (i.e. university faculty / staff / students, general public, etc.) and
- indicate what resources they are allowed access to and/or
- what actions they are allowed to perform on those objects.

Status: Basic development in place; need better integration across systems

Example: Argus/AIMS, Shibboleth

Access Policies

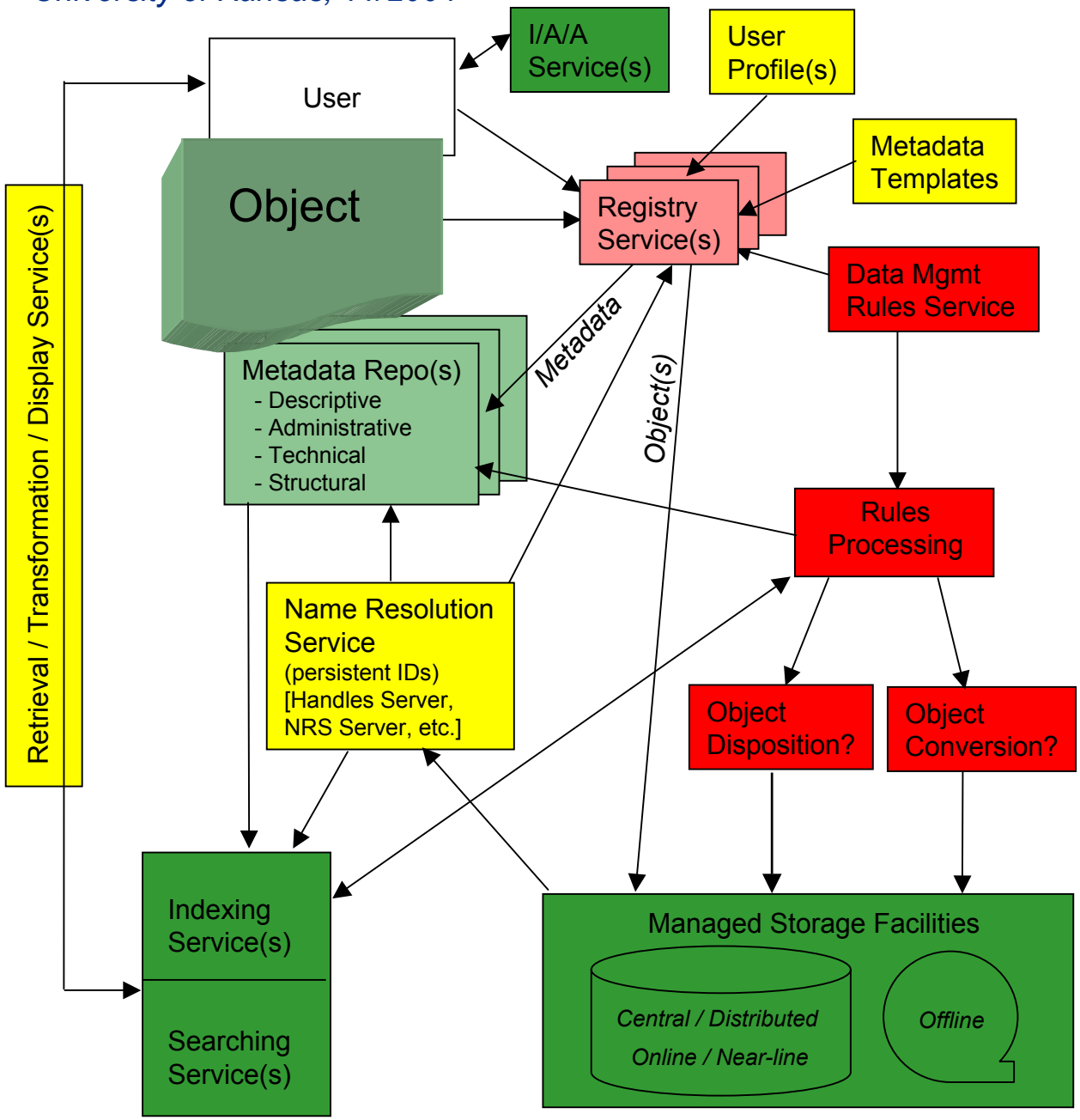
Other Policies

In place (minimal)

Under development

Not yet started

University of Kansas, 11/2004



Digital Objects:

Digital material submitted for registration and management:

- can be physical or logical.
- sufficient associated metadata (supplied by submitter and / or system) for access, retrieval, and management.
- identified via non-contextual URN(s) contained in the object's administrative metadata and linked to physical storage location(s) via *Name Resolution Service(s).*

Status: Exist

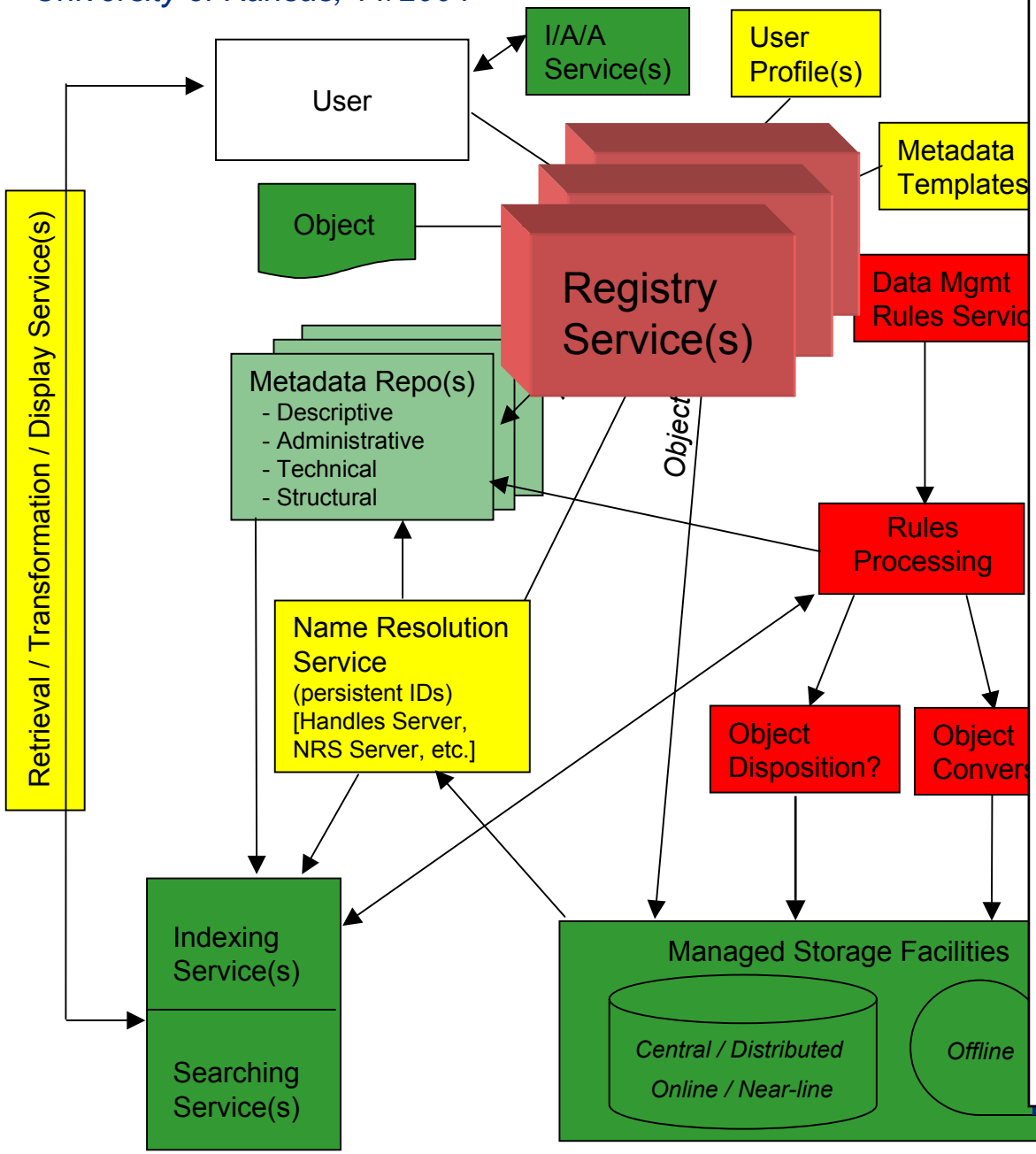
Example: scholarly papers, learning objects, datasets, administrative data, etc.

In place (minimal)

Under development

Not yet started

University of Kansas, 11/2004



Registry Services:

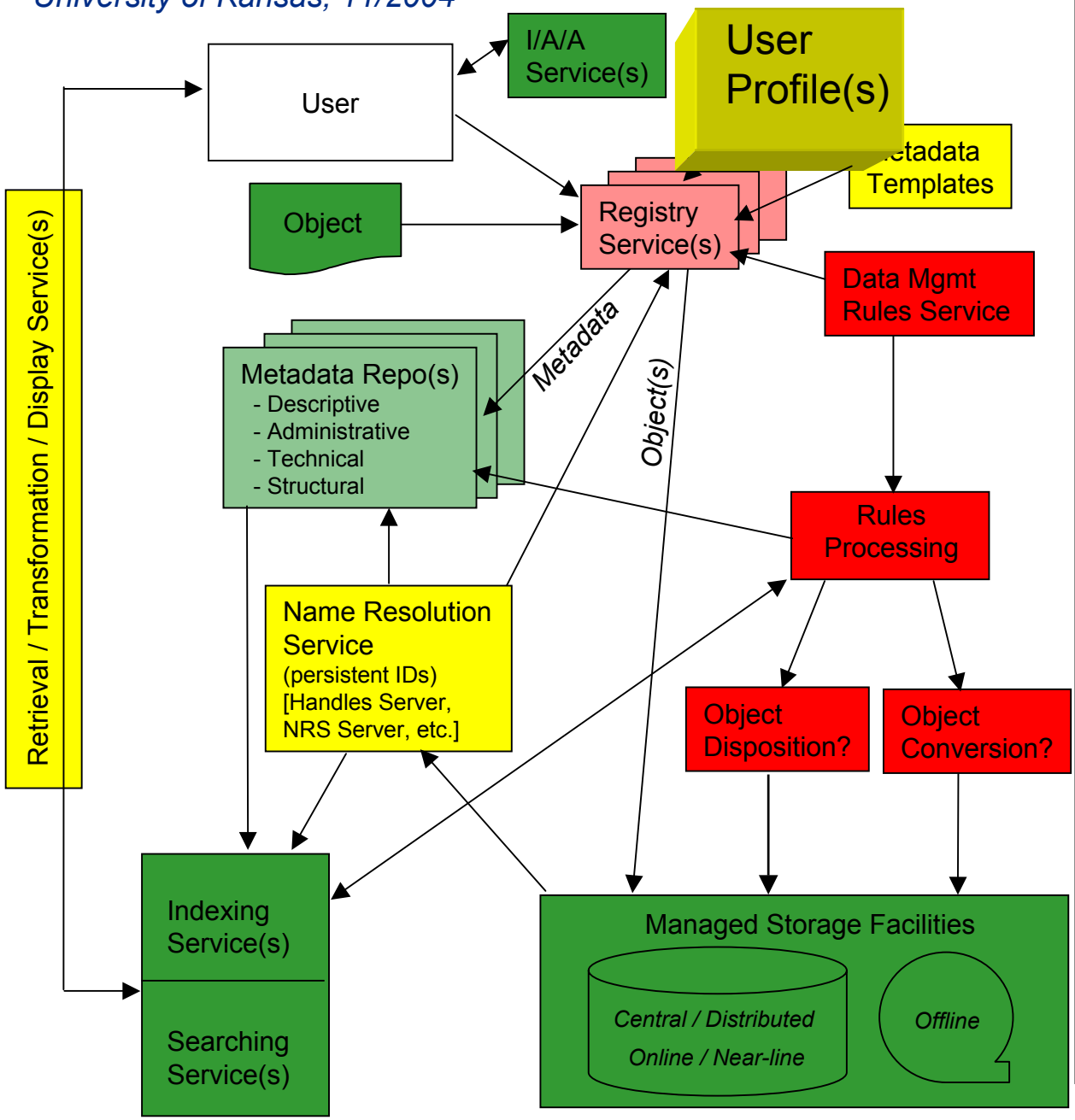
Provide initial ingest and registration of digital objects into overall management infrastructure.

- draw on additional services such as *User Profiles*, *Metadata Templates*, *Data Management Rules,* and *Name Resolution Service(s)*
- lead submitters through processes and decisions required to:
  - collect descriptive, technical, and administrative information
  - apply appropriate classification and disposition elements,
  - determine appropriate metadata & storage repositories, and
  - assign an appropriate URN.
- available for both interactive and batch submissions
- web-based and/or
- include a desktop "drop-box" feature for ease of use

Status: Minimal for selected systems; need generic, standalone service that interacts with multiple systems

Example: Submission process for DSpace (KU ScholarWorks)

Not yet started



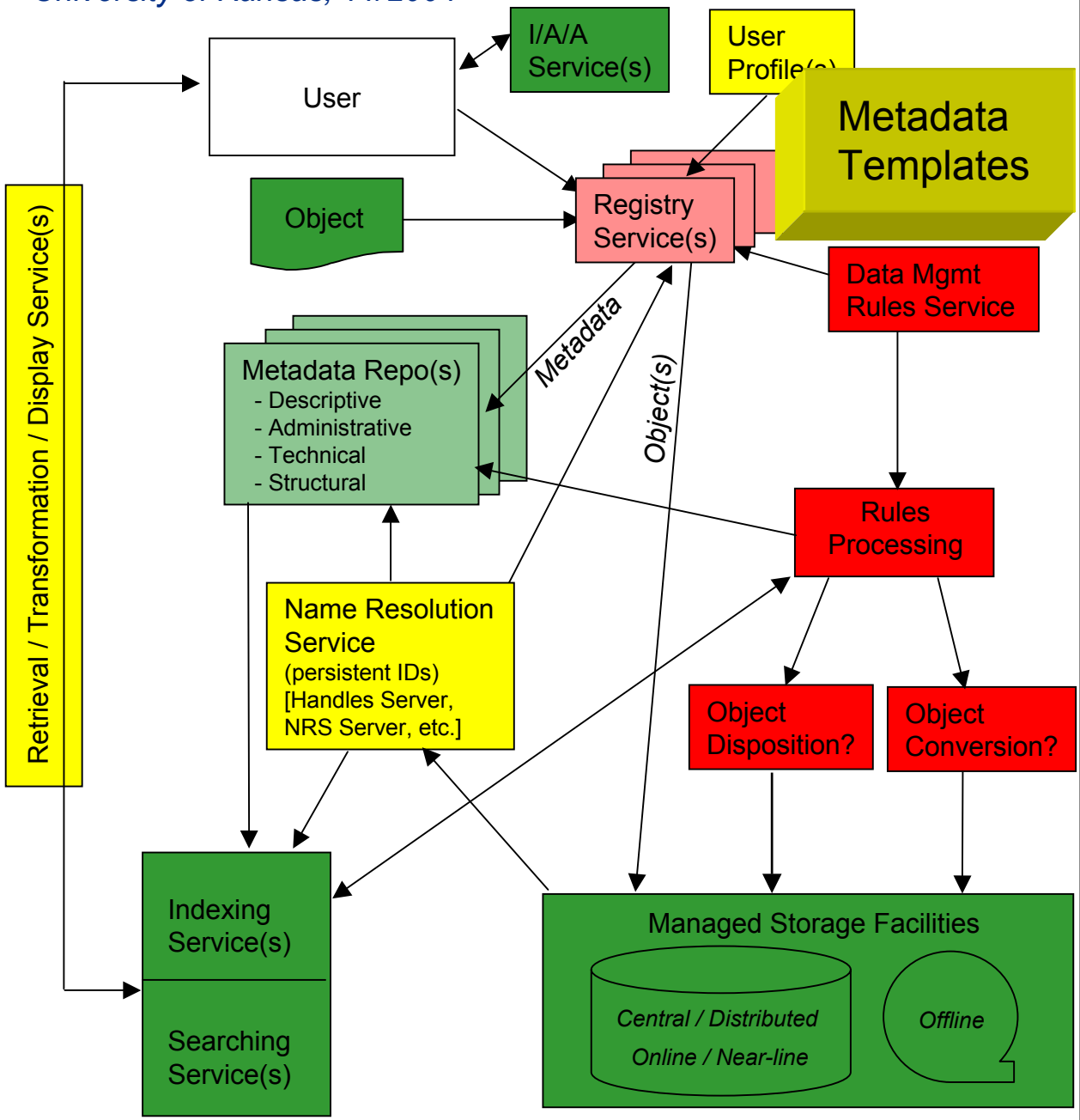
User Profiles:

Include basic information about the object submitter.

- elements could include
  - correct form of name,
  - department affiliation(s),
  - address,
  - default repository(ies),
  - default object classifications / groupings,
  - default subject terms / keywords,
  - default access rights designations, etc.
- linked to authoritative databases such as
  - directory services and authority files for name / address information,
  - *Data Management Rules* for current object classifications / groups,
  - a list of available repositories, etc.

Status: Minimal development

Example: Directory services and authority files for some elements



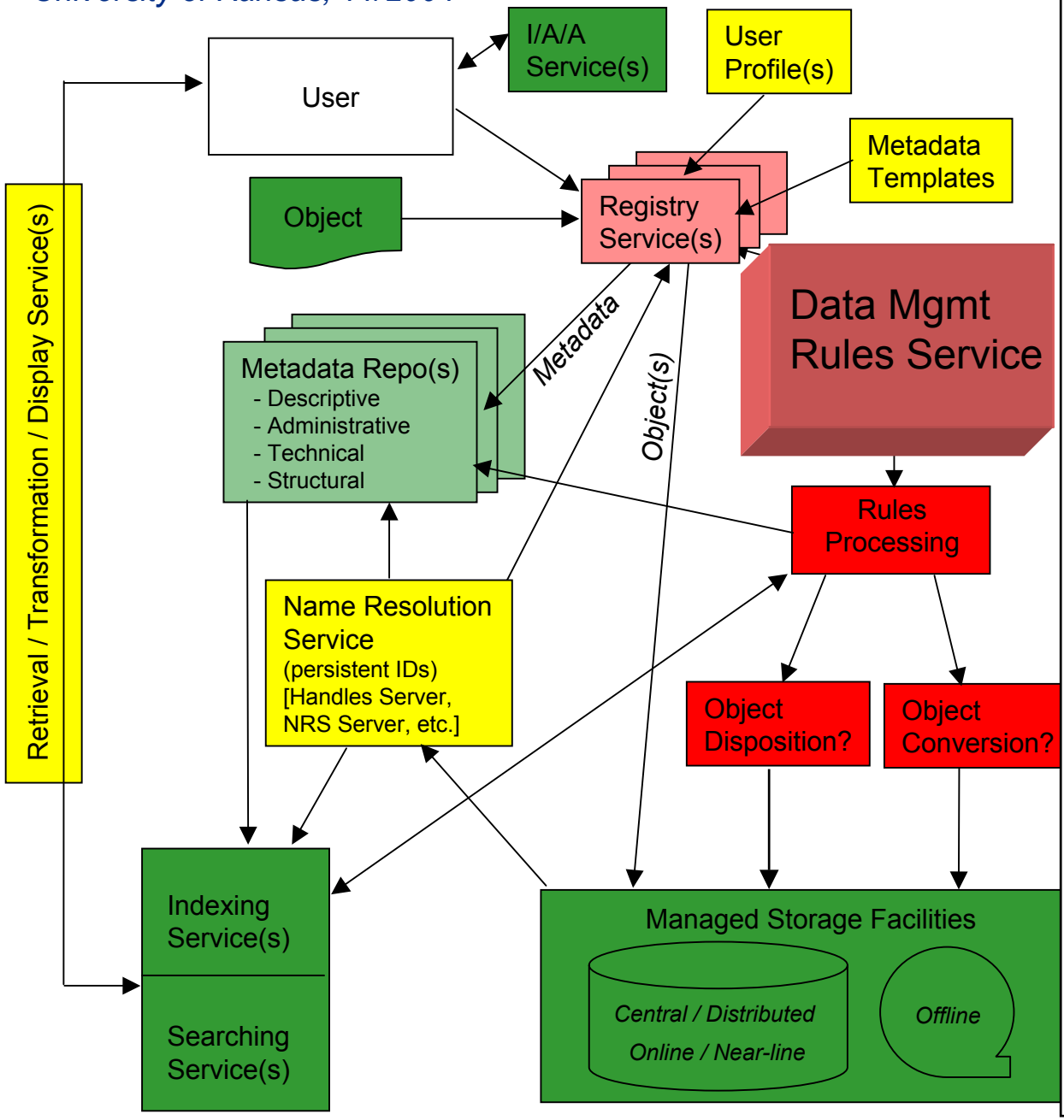
Metadata Templates:

Provide forms for collecting standardized descriptive (Dublin Core, MARC, EAD, IMS, etc.), technical (file format, file relationships, software / version, scanner settings, etc.), and administrative (access rights, classification/disposition, etc.) information.

- Data collected through direct input by the submitter, by system analysis of the submitted object(s), or by system assignment.
- Metadata records are stored in managed *Metadata Repositories*.

Status: In place for selected *descriptive* formats and systems; primarily for staff use; need standalone versions of templates to link to *Registry Service(s)*

Example: Dublin Core format for DSpace, ENCompass; MARC format for Voyager; EAD; Geospatial format for DASC; LUNA



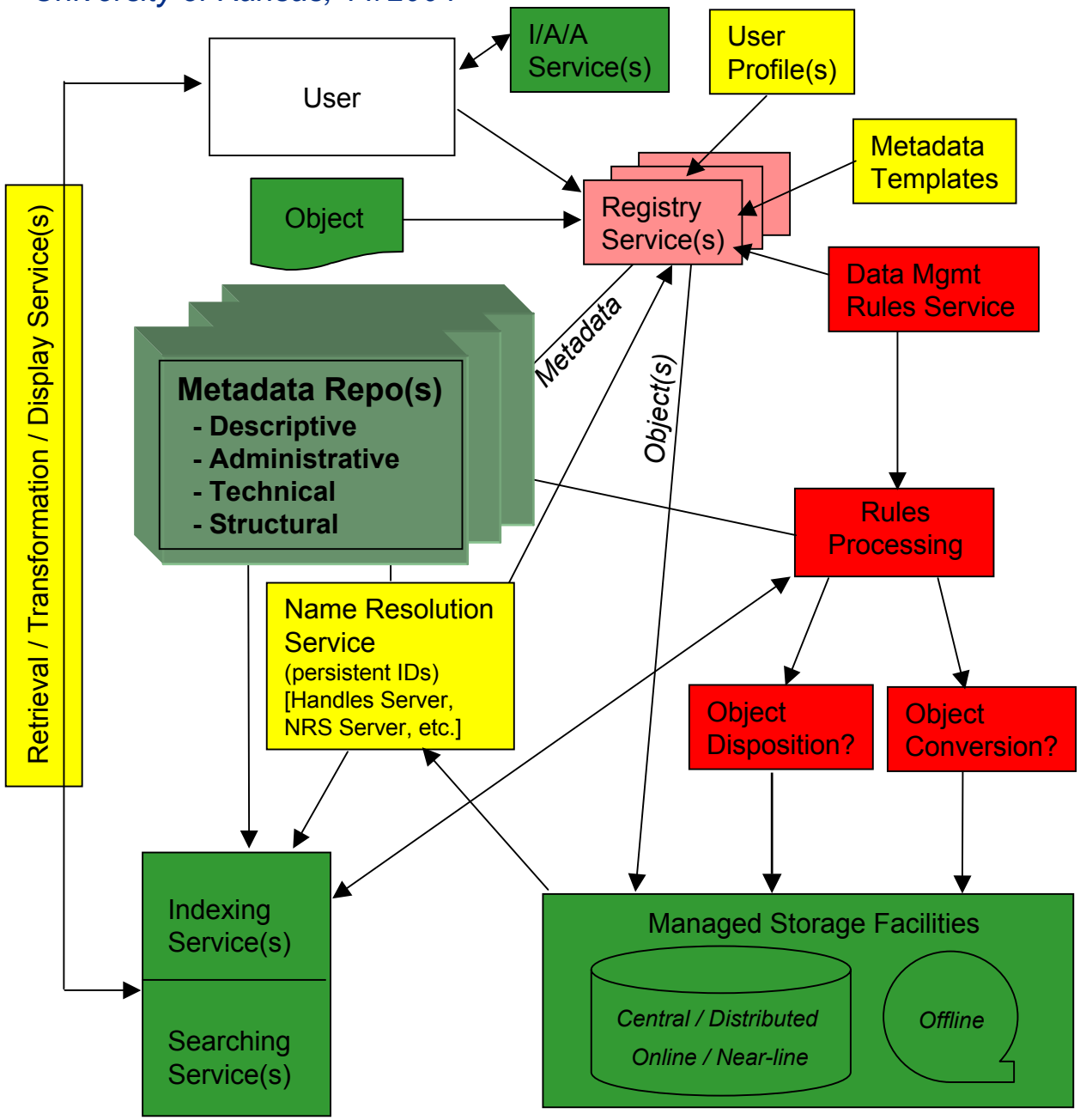
Data Management Rules Repository/Service:

Provides information required for both short- and long-term management of metadata and objects.

- Rules determine such issues as
  - appropriate *Metadata Templates* / metadata formats
  - acceptable object formats,
  - classification / records groups
  - disposition of metadata and objects (methods, dates, etc.),
  - access rights, etc.
- Rules are developed based on approved *standards, policies, and best practices*.
- Rule elements are associated with objects (either directly stored or via pointers, as appropriate) as administrative or technical metadata.

Status: Not yet started locally; need to review for external options

Example: None



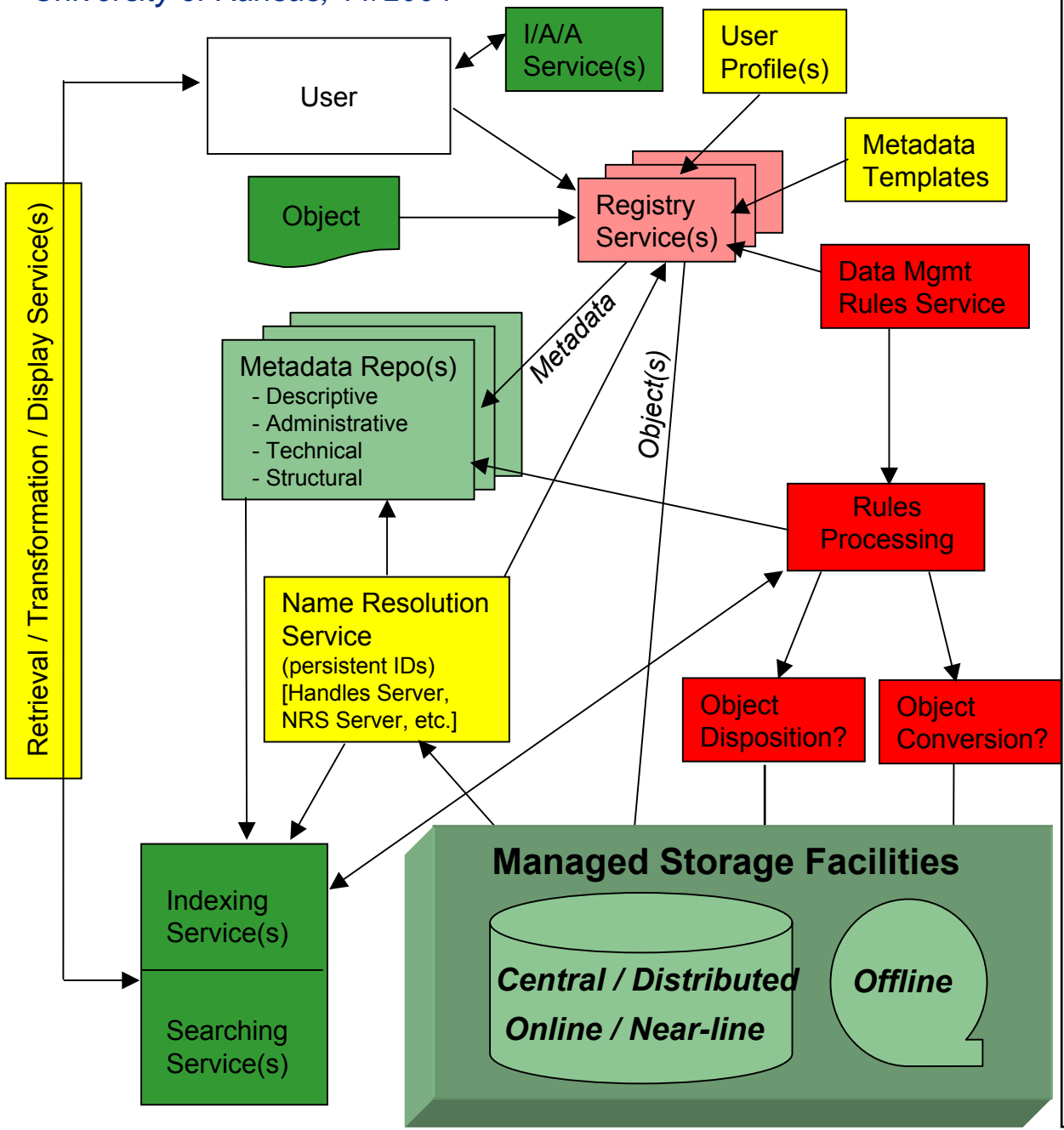
Metadata Repositories:

Managed repositories for storing descriptive, technical / structural, and administrative metadata.

- Managing digital files for ongoing accessibility will be highly dependent on preservation metadata.
- Stored in standardized format such as METS.
- Administrative and/or technical metadata elements will include standardized URNs/URIs linking to objects in object repositories via the *Name Resolution Service(s)*.
- Metadata is maintained (or harvested) and indexed for end-user searching and *Rules Processing* for long-term disposition and/or preservation management of objects.

Status: In place for selected *descriptive* formats and systems; primarily for staff use

Example: DSpace, Voyager, ENCompass, LUNA, PeopleSoft, DASC



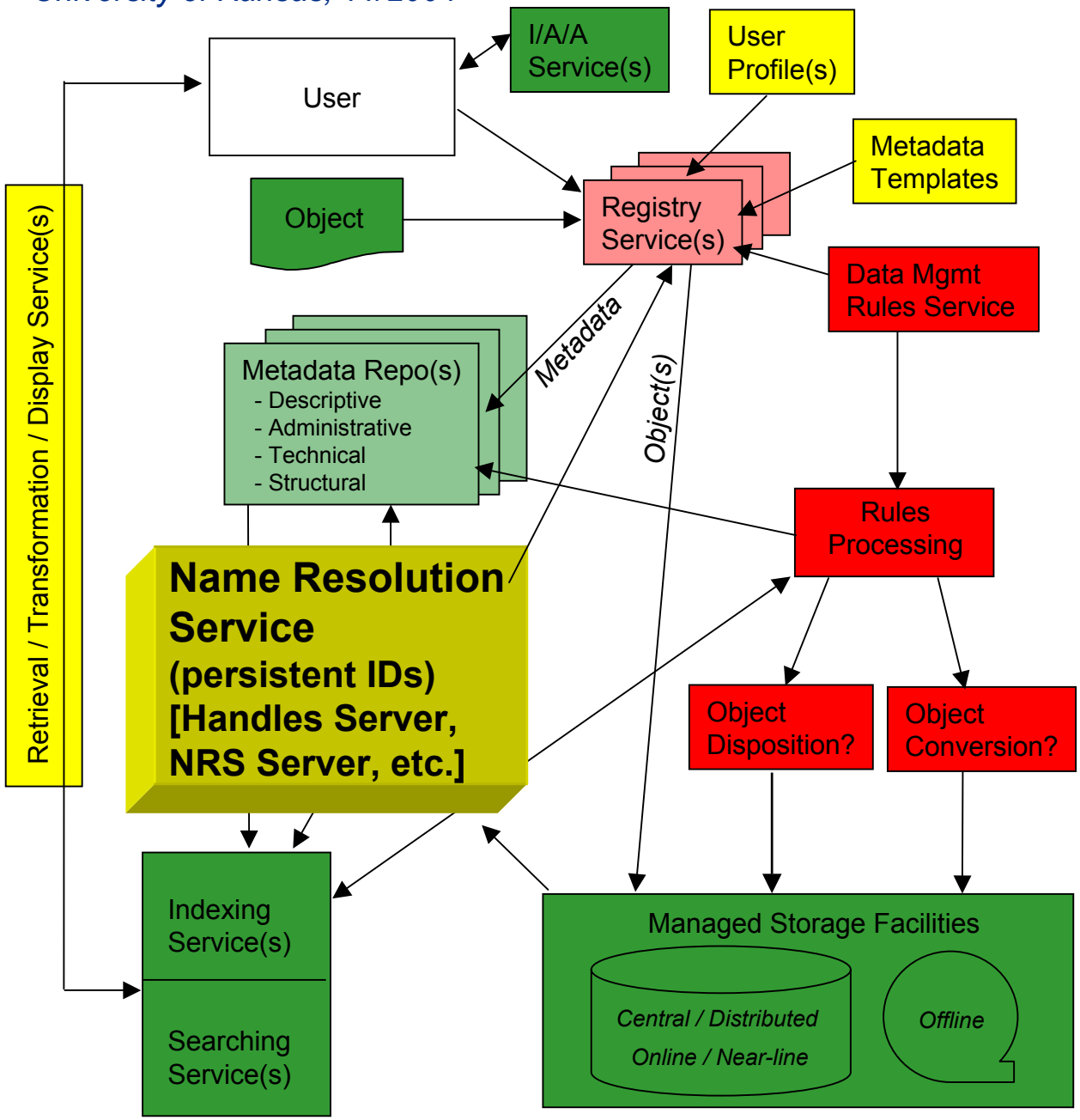
Object Repositories / Managed Storage Facilities:

Managed stores of digital objects.

- repositories can be centrally managed or distributed.
- can be online, near-line, or offline (disk, tape, etc.).
- Objects are maintained as Masters and Access Derivatives
- Objects may include bundled metadata for preservation purposes (self-definition).
- Repositories should be registered and certified based on compliance with approved / standardized storage facility management policies & procedures including backup procedures, refresh / migration management, disaster recovery plans, security procedures, etc.
- File management procedures must tie into *Name Resolution Services* to maintain accurate links to *Metadata Repositories* and records for objects.

Status: In place at a basic level for selected formats and systems

Example: DSpace, Voyager, ENCompass, PeopleSoft, DASC, BEsafe project



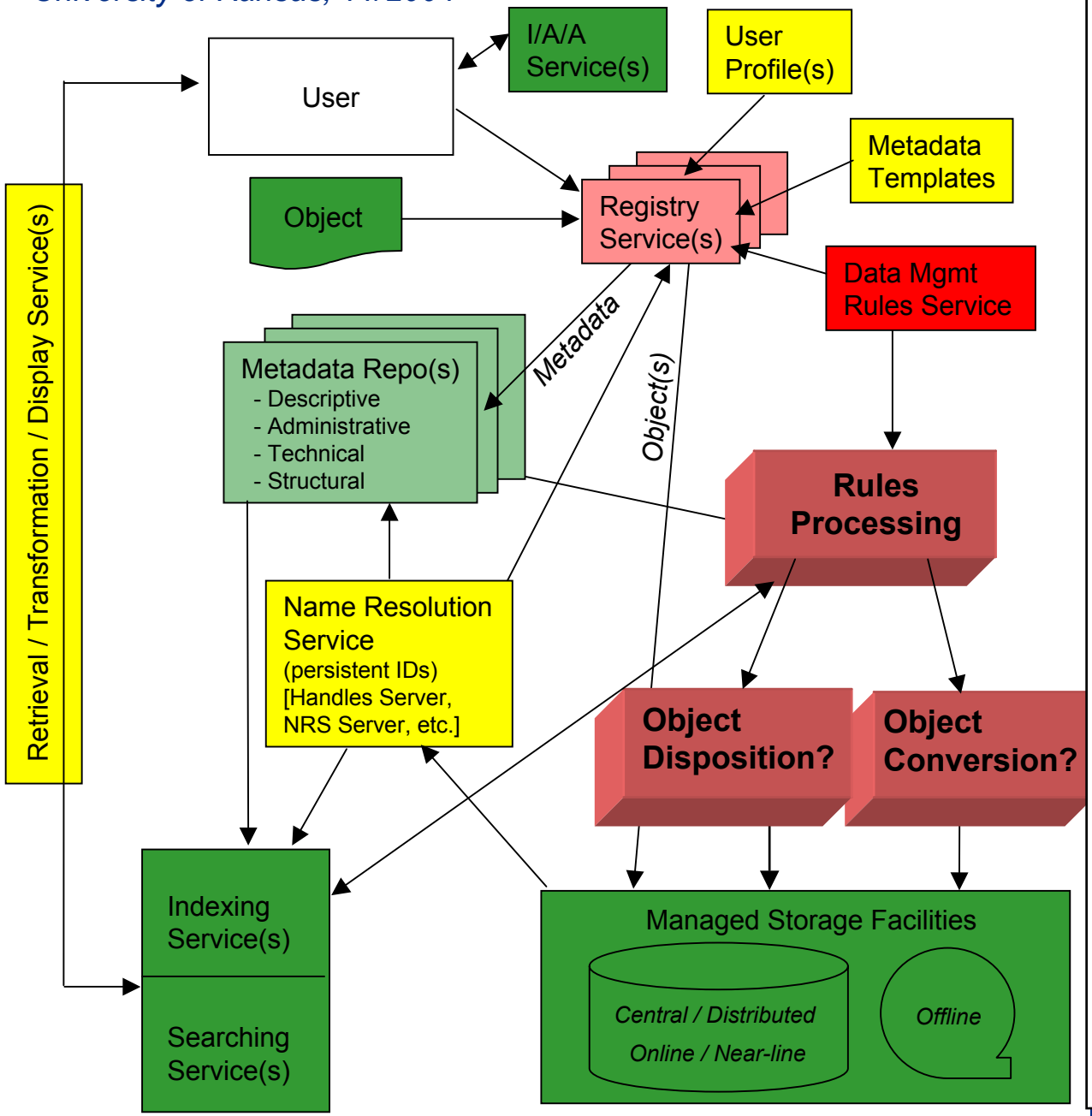
Name Resolution Service(s):

Persistence of links between resources and resource discovery services is essential to ensure long-term access to materials.

- A persistent identifier is an object name (URN: Uniform Resource Name) that remains constant regardless of object location
- Use ensures that when an object is moved, or its ownership changes, links to it will remain actionable.
- Important to note that a *Name Resolution Service* will only be effective if it is maintained. When objects are moved, the current location must be associated with the persistent identifier through use of a resolver database.
- A resolver database is used to translate / map the name (URN) to a current location (URL: Uniform Resource Locator).

Status: Minimal implementation; need decision on option(s)

Example: Handles (CNRI), NRS (Harvard), Purls (OCLC)



Rules Processing:

Processes / utilities that can be invoked for a variety of management tasks such as:

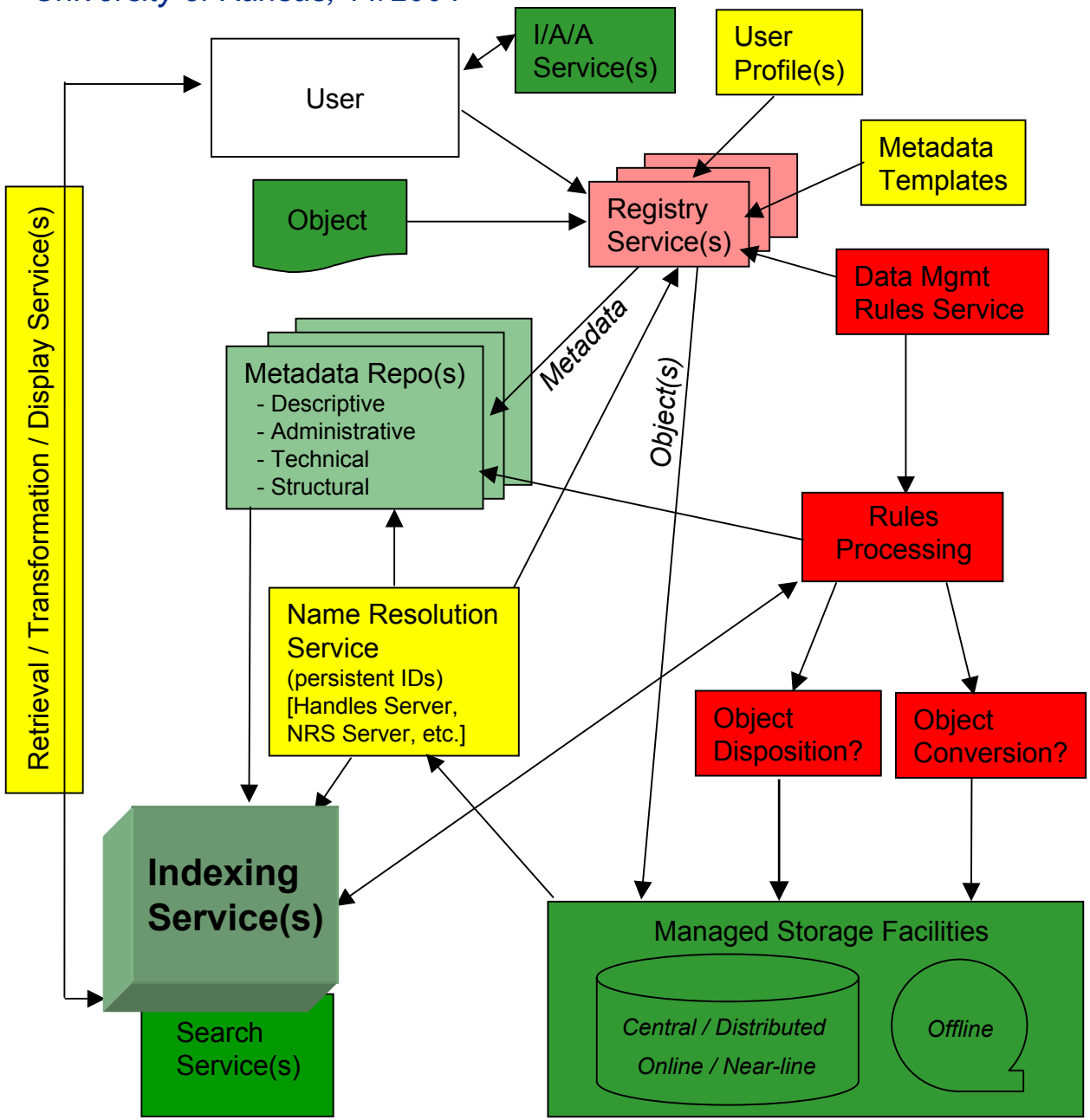
- applying record disposition rules (deleting expired files, changing access rights, etc.)
- identifying object normalization processes / utilities
- identifying objects in specific formats for conversion to function with new software versions
- identifying objects to be migrated from online to offline storage, etc.

- *Rules Processing* must interact with the *Data Management Rules Service*, *Metadata Repositories*, file management utilities, conversion utilities, and *Name Resolution Service*.
- Audit trails detailing actions taken should be maintained

Status: Not yet started locally; need to review external options

Example: None

University of Kansas, 11/2004



Indexing Service(s):

Index descriptive, technical, and administrative metadata and object content (as appropriate) stored in managed, registered *Metadata* and *Object Repositories*.

- These services will generally be distributed and associated with specific applications.
- Object identifiers should be based on URNs; resolution of URNs should be through the *Name Resolution Service(s).*
- Indexes can be used by end-users and by management utilities such as *Rules Processing.*

Status: In place for selected descriptive formats and systems; primarily for staff use

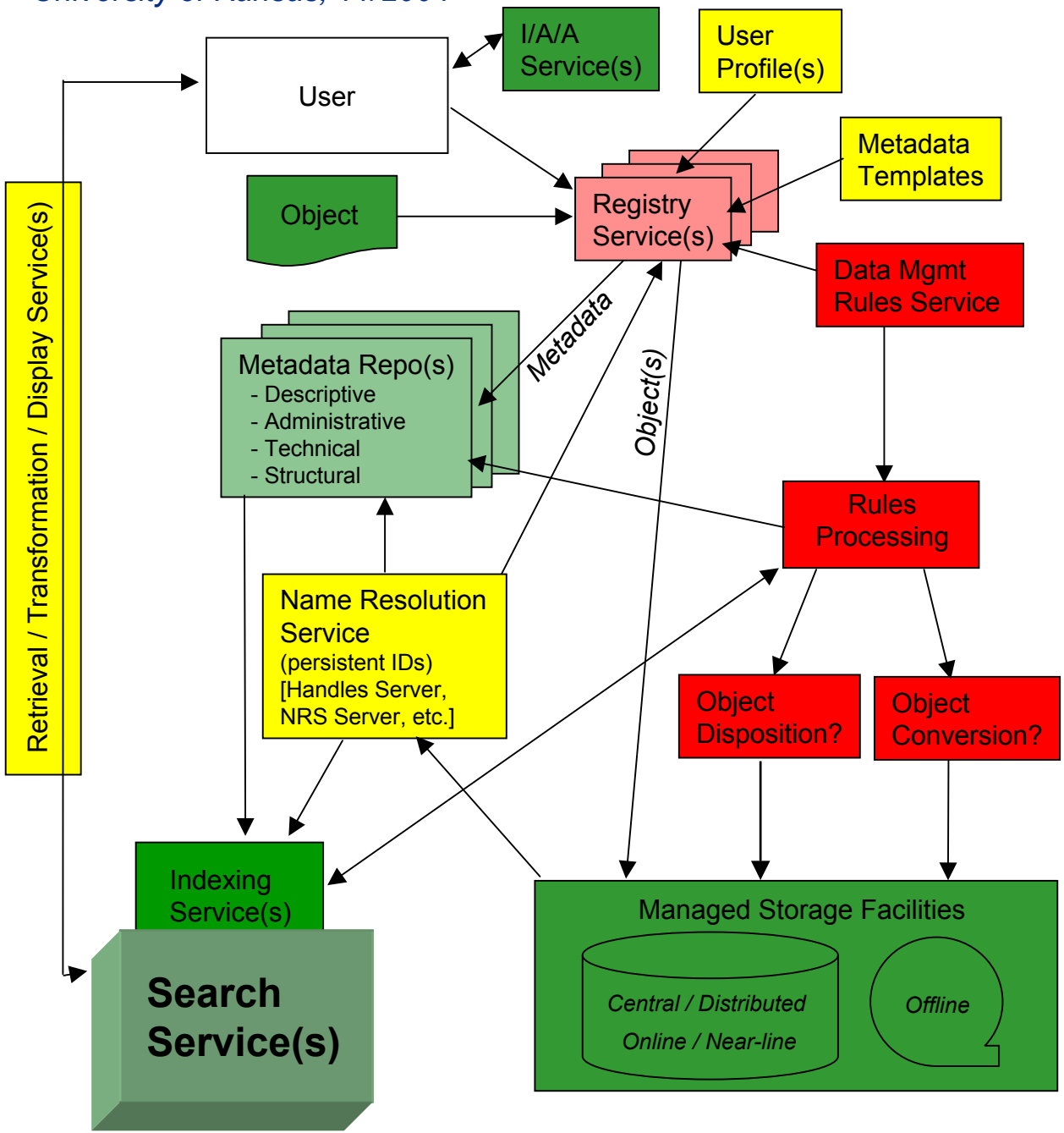
Example: DSpace, Voyager, ENCompass, DASC, campus website Google service

In place (minimal)

Under development

Not yet started

University of Kansas, 11/2004



Search Service(s):

Provide capability to search indexes and discover object metadata and content.

Search services will generally be distributed and associated with specific applications

Status: In place for selected descriptive formats and systems; primarily for staff use

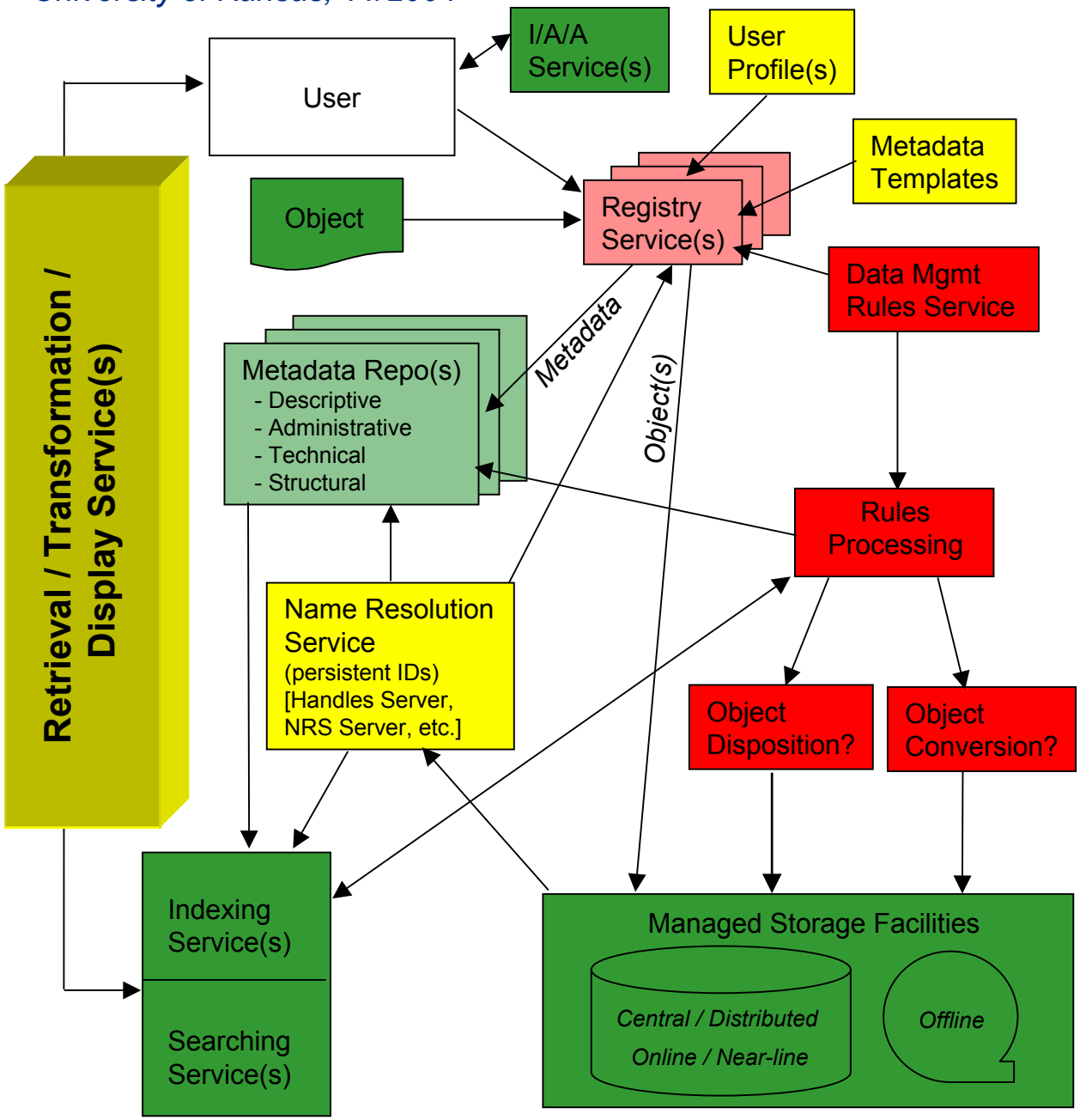
Example: DSpace, Voyager, ENCompass, DASC, campus website Google service

Migration Policies

Access Policies

Other Policies

- *In place (minimal)*
- *Under development*
- *Not yet started*



Retrieval / Transformation / Display Services:

Processes / utilities that

- retrieve metadata and objects from managed repositories;
- dynamically render metadata / objects into usable formats through either transformation or emulation processes; and
- display rendered objects to the user.

These services work in conjunction with *Indexing* and *Search Services*.

Status: some utilities in place for selected transformation services; Web browsers for display

Example: XML to HTML (ENCompass); batch processes for conversion to MrSID and JPEG2000 (GIS); IE, Firefox/Mozilla, Safari

In place (minimal)

Under development

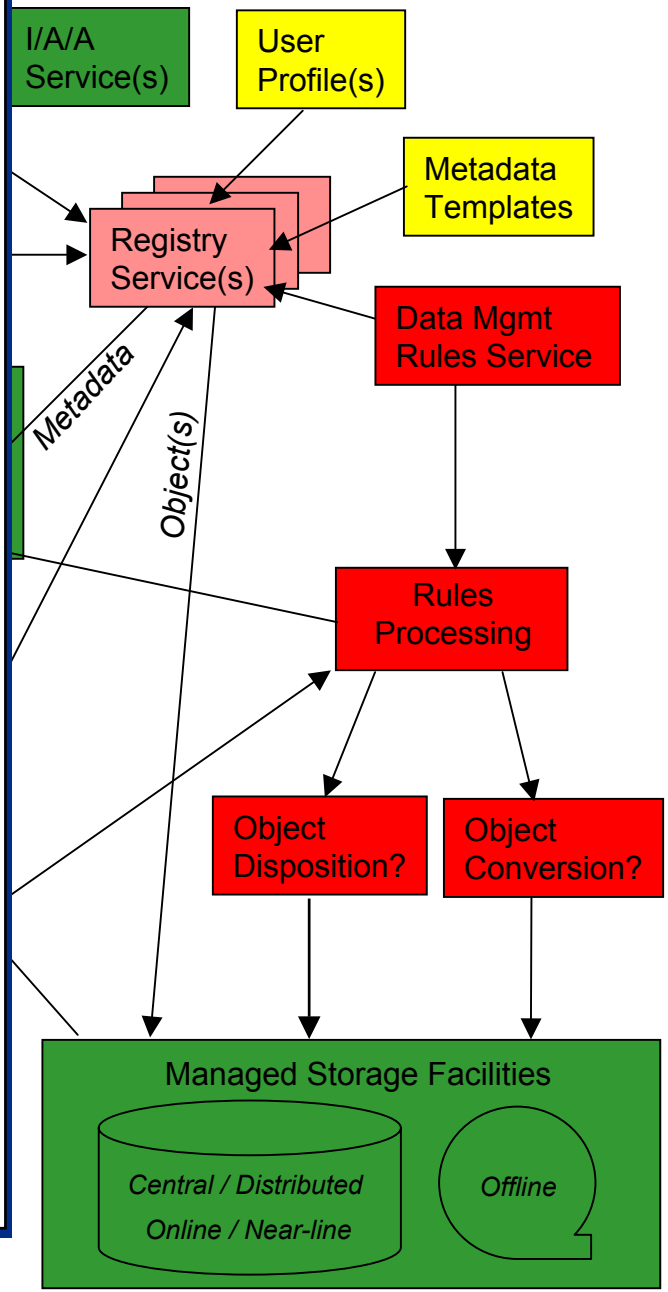
Not yet started

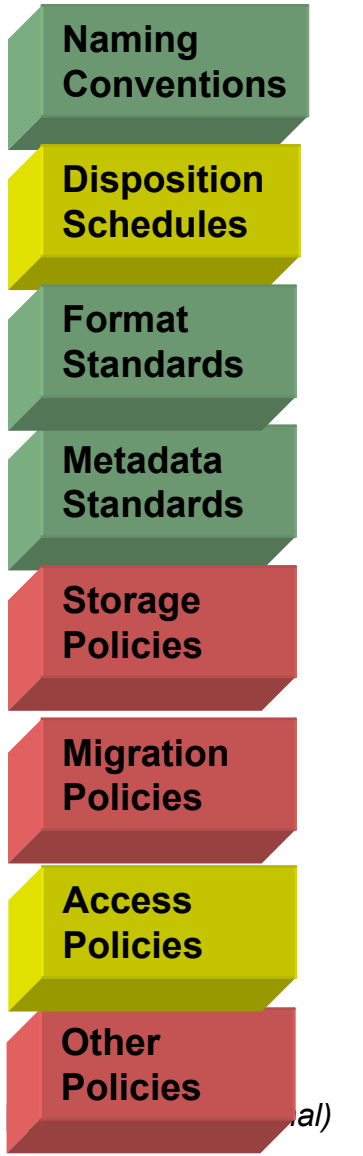
Standards / Policies / Best Practices:

- The use of standards is one strategy used to assist in preserving the integrity of and access to digital information.
- Use of standards can also help ensure best practice in the management of digital information.
- There are standards for many aspects of storing and accessing digital information, including interoperability, data formats, resource identification, resource description, data archiving, and records management.
- Digital preservation policies give structure and general direction for specific actions.

Status: Under development, various stages

Example: Digital Initiatives recommended standards; DASC standards; DSpace policies

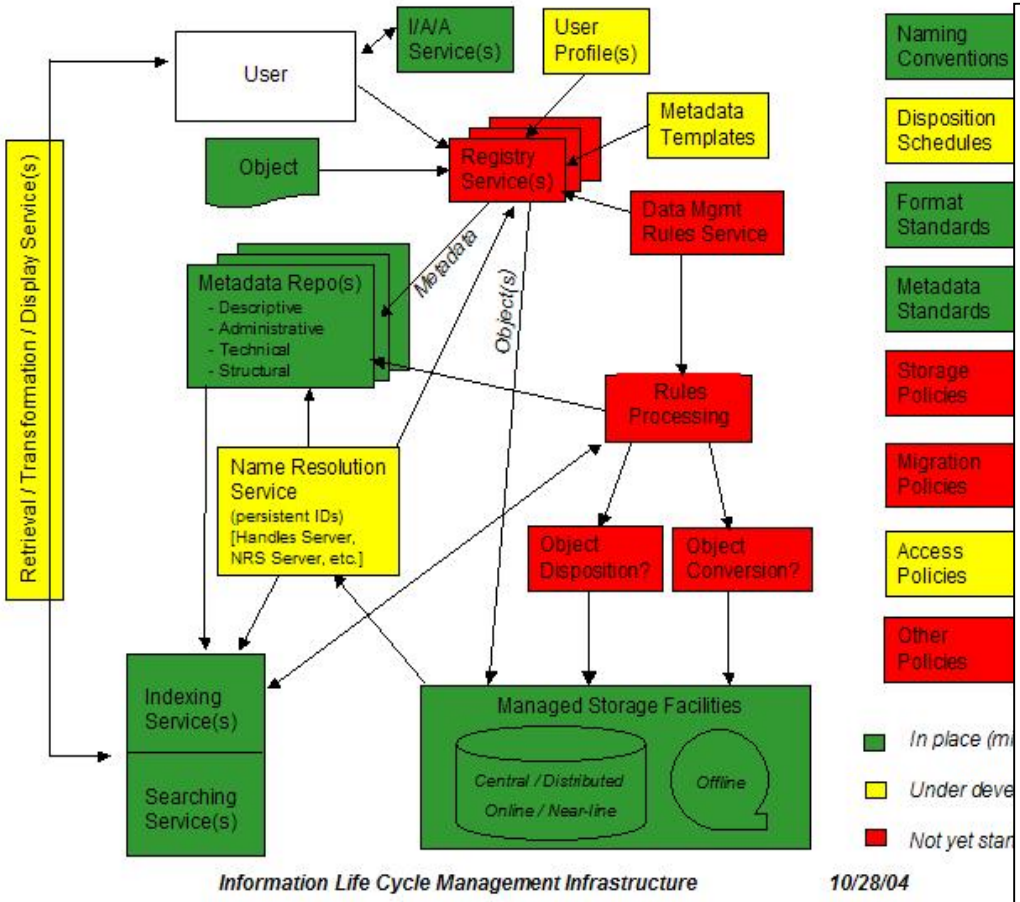




Under development

Not yet started

Technical Infrastructure: Recommendations for Development

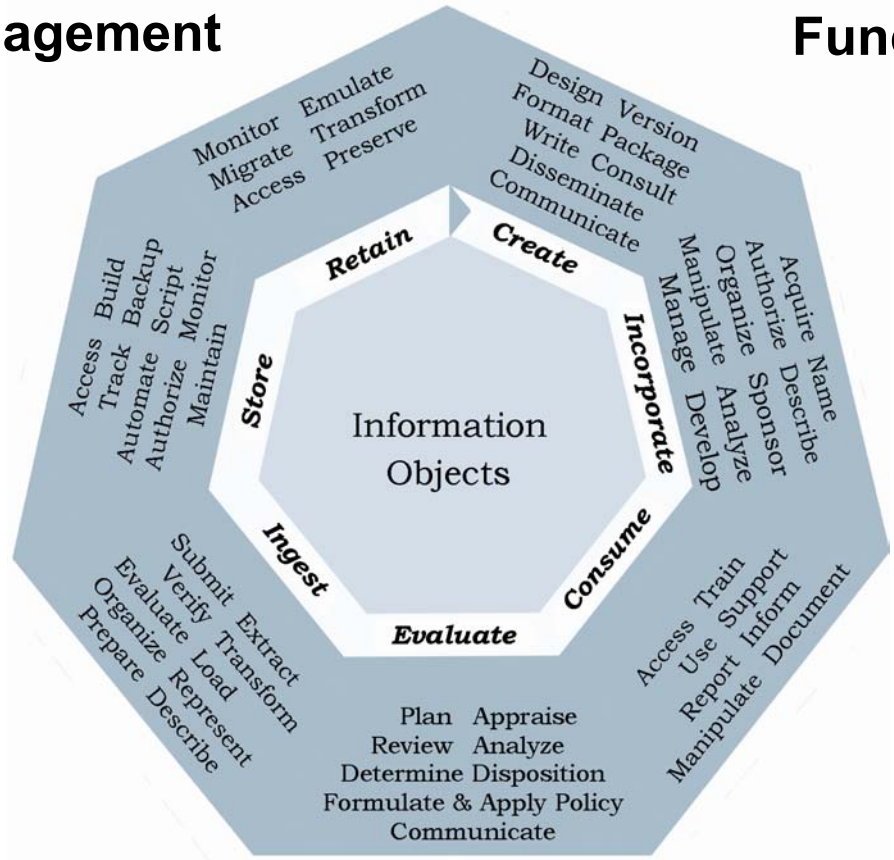


- **Review** existing major digital asset systems
- Develop specifications for and begin development or purchase of components that are currently missing.
  - **Initial work** should be focused on *Registry Services, Data Management Rules Service, and Name Resolution Services.*
- **Identify responsible units** for management of specific ILM components

Organization: Recommendations for Strengthened Roles

Lifecycle Management

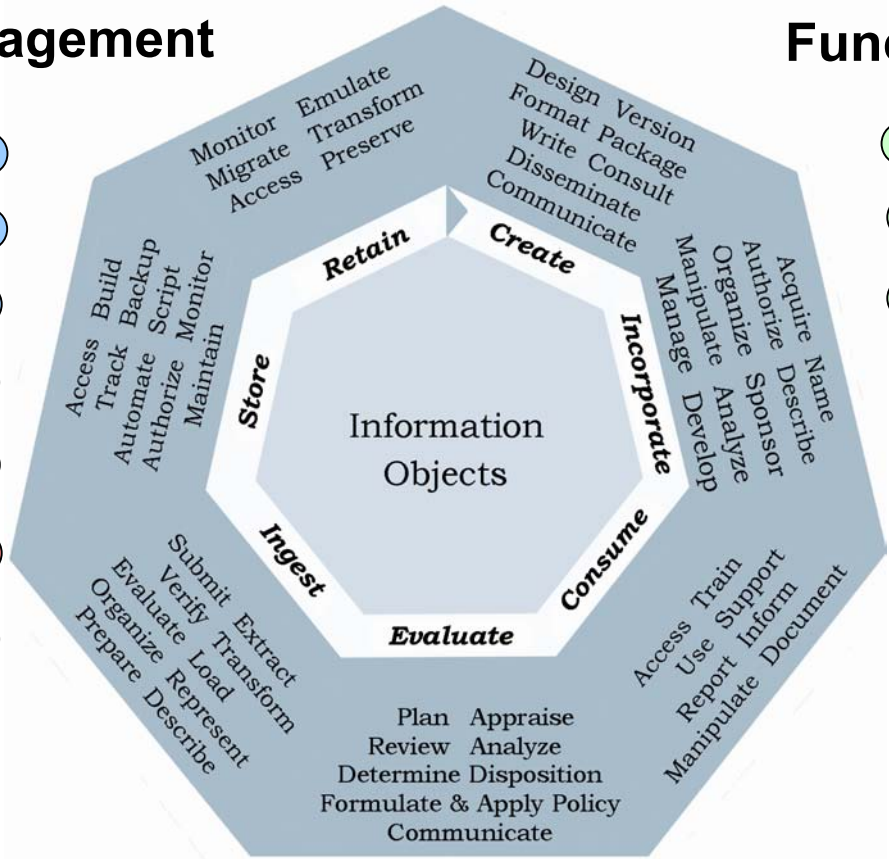
Functional View



Organization: Recommendations for Strengthened Roles

Lifecycle Management

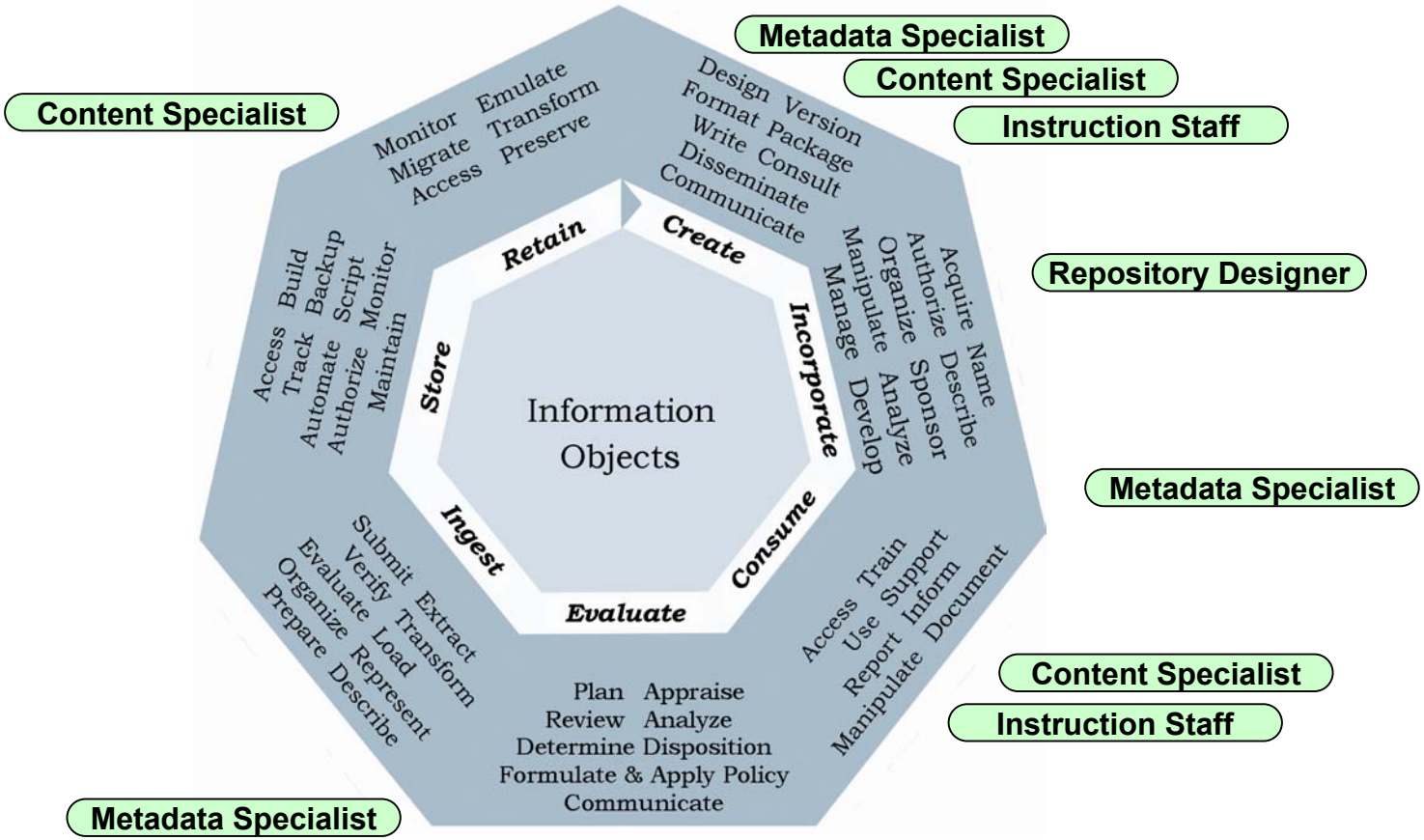
- IT Operations Staff
- Application Specialist
- Data Analyst
- Programmers
- Database Developer
- Policy Analyst
- Policy Administrator



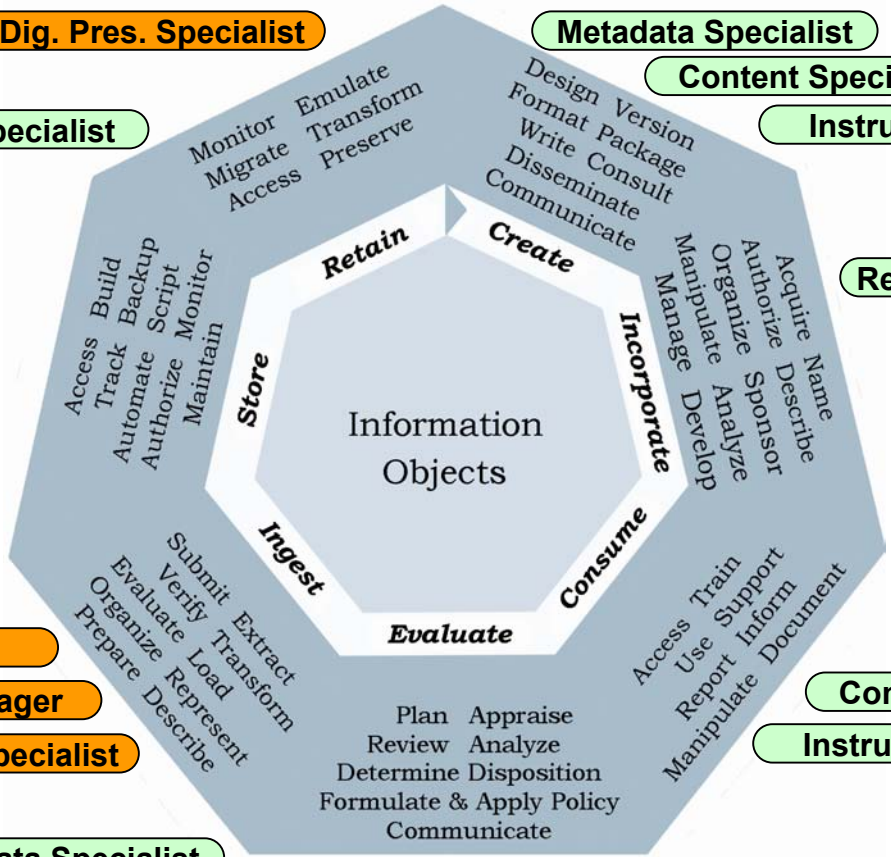
Functional View

- Metadata Specialist
- Content Specialist
- Repository Designer
- Instruction Staff
- Archivist
- Records Manager
- Dig. Pres. Specialist
- Dig. Pres. Officer

Organization: Recommendations for Strengthened Roles



Organization: Recommendations for Strengthened Roles



Dig. Pres. Specialist

Metadata Specialist

Content Specialist

Content Specialist

Instruction Staff

Repository Designer

Metadata Specialist

Archivist

Records Manager

Dig. Pres. Specialist

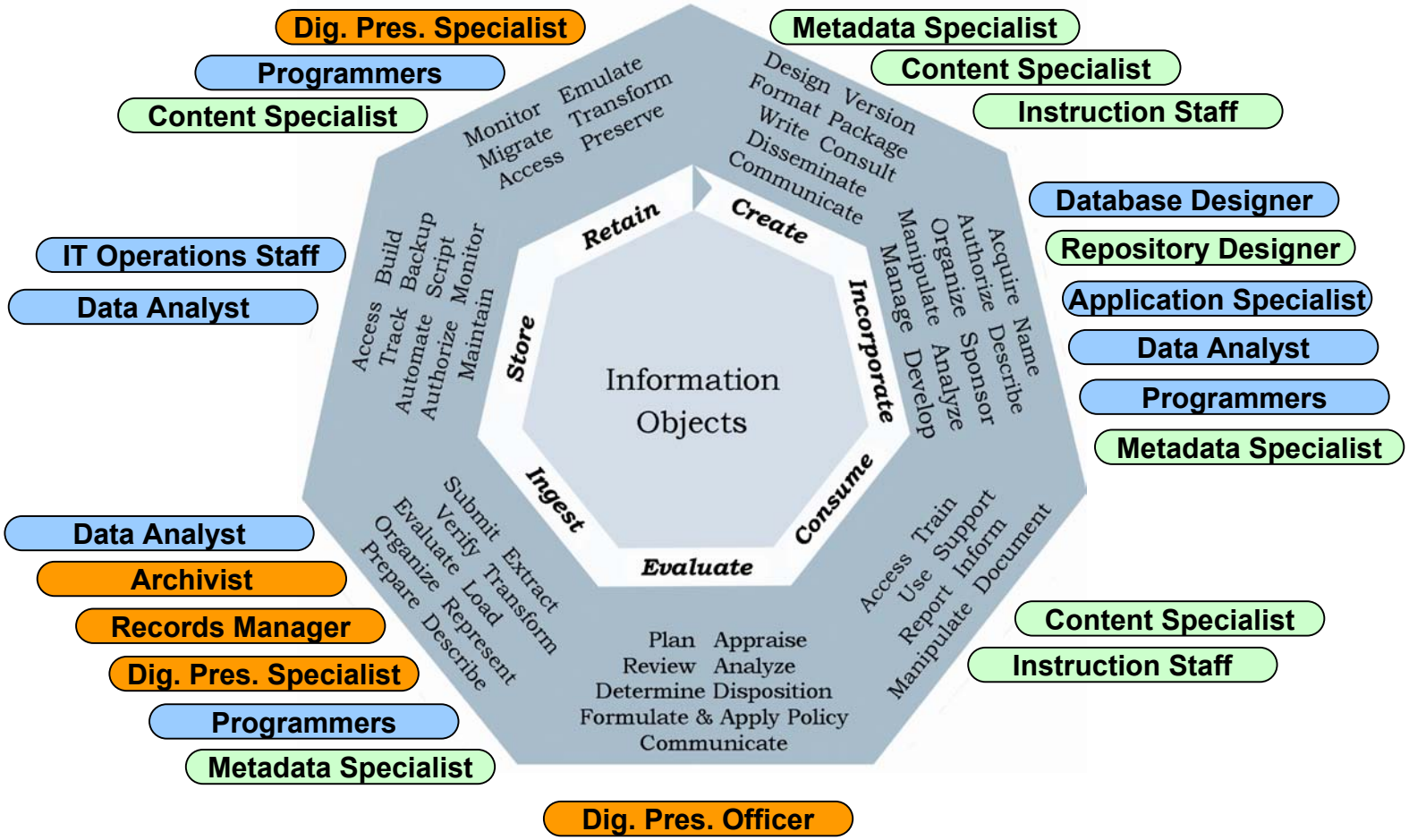
Content Specialist

Instruction Staff

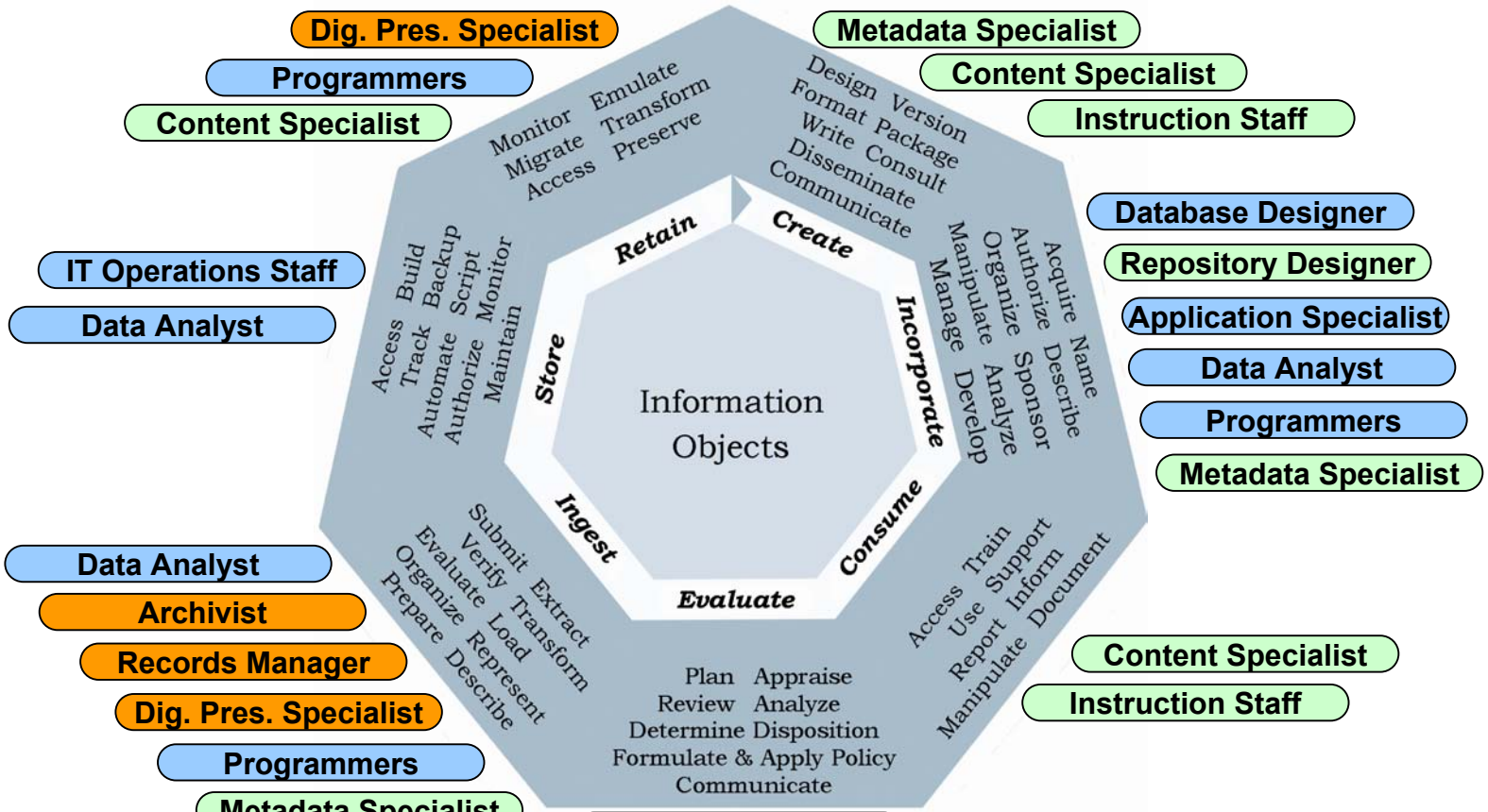
Metadata Specialist

Dig. Pres. Officer

Organization: Recommendations for Strengthened Roles



Organization: Recommendations for Strengthened Roles



Organization: Recommended Policies for Completion or Development

- **Policies for Appraisal:** Not all university digital assets can be preserved; appraisal policies need to be developed to set priorities.
  - Administrative record appraisal and schedules
  - Non-administrative digital assets held by University units
    - Develop a template by which units can identify and inventory classes of digital assets under their stewardship
    - Develop guidelines to help determine their relative priority for long-term retention
- **Policies for Creation of New Content**
  - Recommended file formats
  - Intellectual property for academic work by KU authors and deposit of academic work in KU ScholarWorks
    - University or governance resolution calling upon all faculty to
      - deposit scholarly work of enduring value in the KU ScholarWorks repository and
      - attempt to retain certain rights in their published scholarship, including the right to disseminate it through a university repository.
  - Permission to copy work on KU servers for preservation administration
- **Policies for Resources and Infrastructure**
  - Best practices for Technical Liaisons and System Administrators outside the Information Services structure
  - Resource allocation

Outreach: Recommended Education Program

A Digital Preservation Curriculum for University Staff

This draft curriculum outlines both broad and specific topics that should be covered in education and training for different groups of university staff. It is recommended that IS and other staff involved in the creation and support of digital information participate in the development, implementation and assessment of the specific training modules.

Outreach: Recommended Education Program

Module I. General Awareness for all University Staff

GOAL: To introduce the concepts of digital preservation at the University of Kansas

- What is Digital Preservation?
- Why is Digital Preservation Important?
- Who is sponsoring the Digital Preservation movement at KU?
- What is the role of Information Services (Libraries and IT) in Digital Preservation?
- What is your role in Digital Preservation?
  - As creator
  - As office manager
  - As researcher or end-user
- What resources are available for further learning?

Outreach: Recommended Education Program

Module II. Life Cycle Management

GOAL: To make staff aware of lifecycle management decisions that impact digital preservation.

- The Lifecycle of Information Management
  - Introduction to the Open Archival Information Systems (OAIS) model
  - Lifecycle Management Issues
    - Selection, Migration, Object integrity, Authenticity, Emulation, Software tools, Hardware issues
  - Stewardship issues (responsibility, ownership, cost)
- Key Issues When Acquiring New Systems
  - Standards adherence, Object formats, Metadata capabilities
  - Ability to export / import data, OAI support
  - Access and Authentication / authorization
  - Availability of information to external search services (federated search, Google)
  - Digital rights management (DRM)

Outreach: Recommended Education Program

Module III. Storage Management & Repositories

GOAL: To discuss safe storage of data, differences between storage and preservation and options for making data accessible while encouraging preservation

- Storage Systems - Where / How to Archive Your Data
  - Central vs. distributed
  - Databases vs. file systems
  - Backup formats, media, schedules, compression
  - Documentation
- Repository Decisions
  - Repository choices for best use of information
  - Long-term preservation of digital masters
  - Data migration pathways for longevity

Outreach: Recommended Education Program

Module IV. Standards

GOAL: To discuss standards adherence and the role of standards in making sure that important information can be preserved.

- Types of information:
  - Administrative Data & Electronic Records
  - Images and Media
  - Documents & Marked Text
  - Databases and Datasets
  - Numeric, GeoSpatial, and Other Interesting Kinds of Collections
- Metadata standards and their application
- Interoperability standards and considerations

Outreach: Recommended Education Program

Module V. Legal Issues

GOAL: To review legal issues to be considered in preserving digital information.

- Copyright / DMCA
- Privacy (HIPPA, FERPA)
- Records management (retention, access control, etc.: state, national)
- Granting agency policies (NIH, NSF, etc.)

Summary of Recommendations

- Technical Infrastructure
- Strengthen Selected Staff Roles
- Develop or Complete Selected Policies
- Educational Outreach

Not Just "Bit Management"… Architecture for Preservation = Architecture for Digital Scholarship

- Registry services and object repositories
- Metadata templates and repositories
- IAA and Name Resolution Services
- Data management rules and processing
- Organizational roles, policies, and education
- Scholarly sharing
- Discovery and access
- Flexible re-purposing between teaching and research

For Further Information

- Presentation: http://kuscholarworks.ku.edu
- HVC^2 Website: http://www.ku.edu/~hvc2/
- Information Services Website: http://informationservices.ku.edu

Contact Us:

- Richard Fyffe: rfyffe@ku.edu
- Deborah Ludwig: dludwig@ku.edu
- Beth Forrest Warner: bwarner@ku.edu