Perspectives on Digital Preservation

Kenneth Thibodeau, Director
Electronic Records Archives Program
National Archives and Records Administration (NARA)

NARA's Vision for the Electronic Records Archives

"The Electronic Records Archives will authentically preserve and provide access to any kind of electronic record, free from dependency on any specific hardware or software, enabling NARA to carry out its mission into the future."

NARA's Electronic Records Challenge

- Take any type of record

- Created using any type of application

- On any computing platform

- From any entity in the Federal Government, and

- Provide discovery and delivery to anyone with an interest and legal right of access

- Now and for the life of the republic

Managing and Enabling Worlds of Knowledge

"Strategies to assure long-term preservation of digital records constitute another particularly pressing issue for research...."

Networking and Information Technology Research and Development

Supplement to the President's Budget for FY 2002

Advanced Technologies to Build Knowledge from Data

- Ultra-large-scale data-mining technologies
- for rapid mining, filtering, correlating, and assessing of vast quantities of heterogeneous and unstructured data ....

Data storage and management technologies:

- Tools for collecting, indexing, archiving, and synthesis
- Protocols for data compatibility, conversion, interoperability, and interpretation in networked environments ....
- Representation, preservation, and storage of multimedia collections

Purpose of Preservation:

Authentic Preserved Records

What comes out of preservation is identical to what went in, in all essential respects
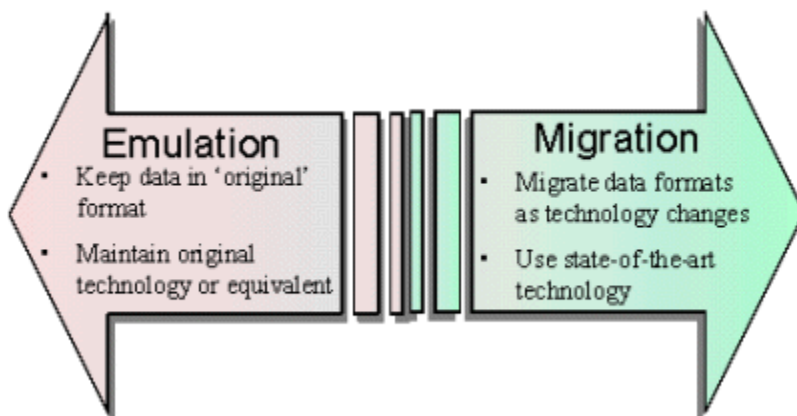
$$X' = X \pm \delta$$

Authenticity

- Is a function of the identification and characterization of the object deemed authentic.
- E.g., what is an authentic copy of Thomas Jefferson's writing?
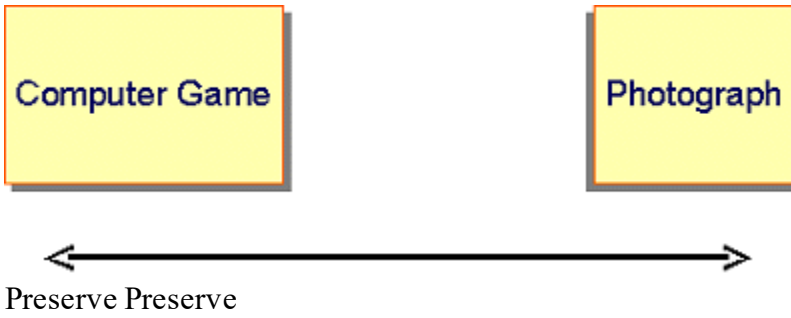-

"The people, especially when moderately instructed, are the only safe, because the only honest, depositories of the public rights, and should therefore be introduced into the administration of them in every function to which they are sufficient; they will err sometimes and accidentally, but never designedly, and with a systematic and persevering purpose of overthrowing the free principles of the government"
Thomas Jefferson to M. Coray, 1823.
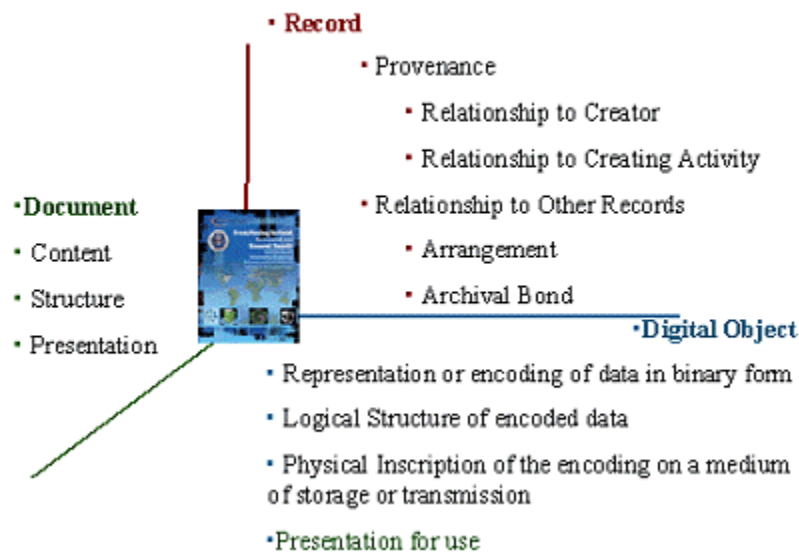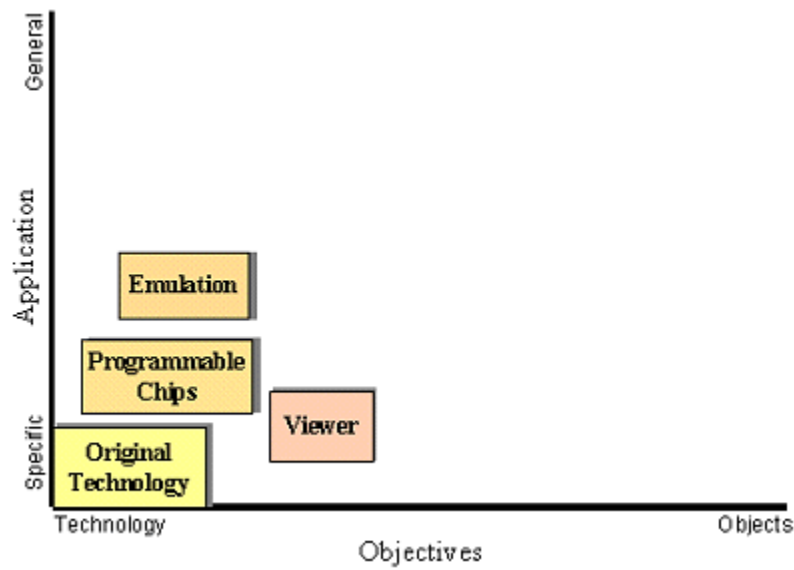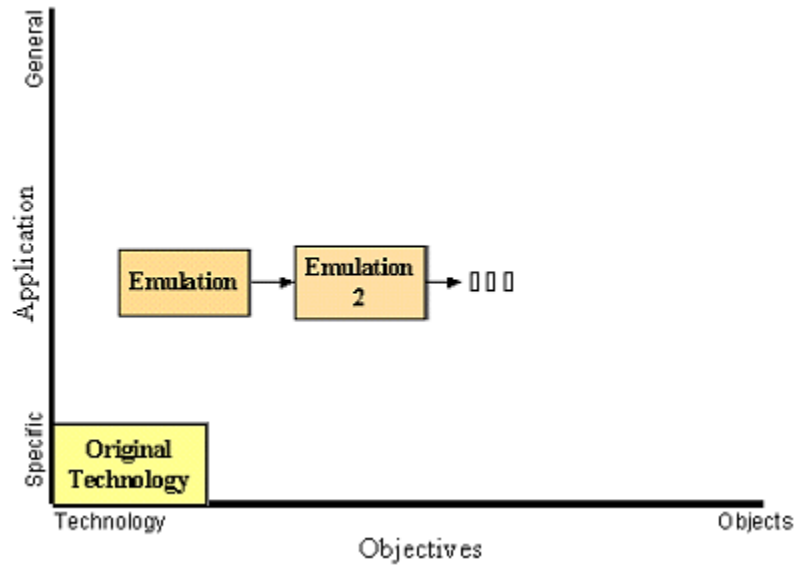ME 15:483

Possibilities for Digital Preservation

Emulation
- Keep data in 'original' format
- Maintain original technology or equivalent

Migration
- Migrate data formats as technology changes
- Use state-of-the-art technology

Possibilities for Digital Preservation

Computer Game   Photograph

Preserve Preserve

Technology Objects

Characteristics of an Electronic Record

- **Record**
  - Provenance
    - Relationship to Creator
    - Relationship to Creating Activity
  - Relationship to Other Records
    - Arrangement
    - Archival Bond

- **Document**
- Content
- Structure
- Presentation

- **Digital Object**
- Representation or encoding of data in binary form
- Logical Structure of encoded data
- Physical Inscription of the encoding on a medium of storage or transmission
- Presentation for use

Chart 1: Axes — vertical "Application" with "General" (top) and "Specific" (bottom); horizontal "Objectives" with "Technology" (left) and "Objects" (right). Boxes: "Original Technology" (bottom left), "Emulation" → "Emulation 2" → □ □ □

Chart 2: Axes — vertical "Application" with "General" (top) and "Specific" (bottom); horizontal "Objectives" with "Technology" (left) and "Objects" (right). Boxes: "Emulation", "Programmable Chips", "Viewer", "Original Technology"
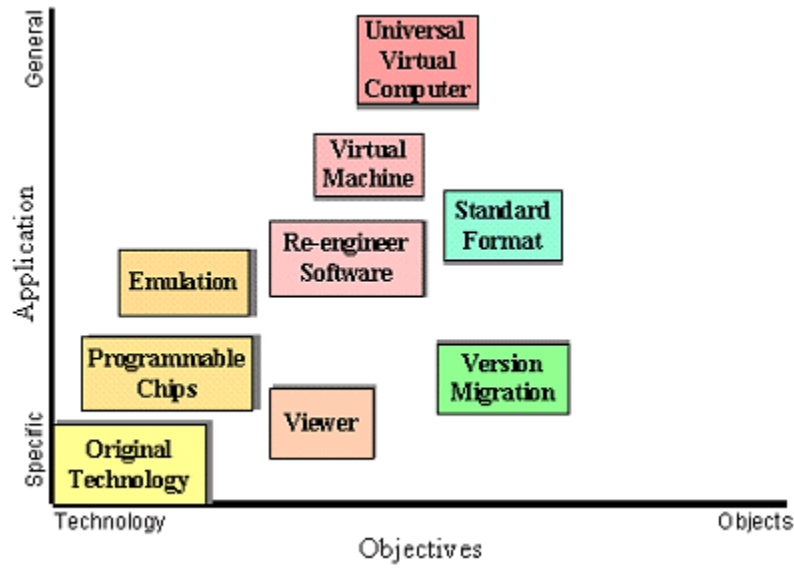
Technology Preservation Methods

- Perpetuate the problems of current technologies
- Increasingly complex to support and use
- Thousands of combinations of hardware, operating systems, application software packages
- Increasingly difficult to train and assist users

- Even more complex as the focus shifts from individual systems to networks, distributed applications, and grid computing

- Becomes very complex for collections of documents accumulated over time

- Run counter to the major direction of IT
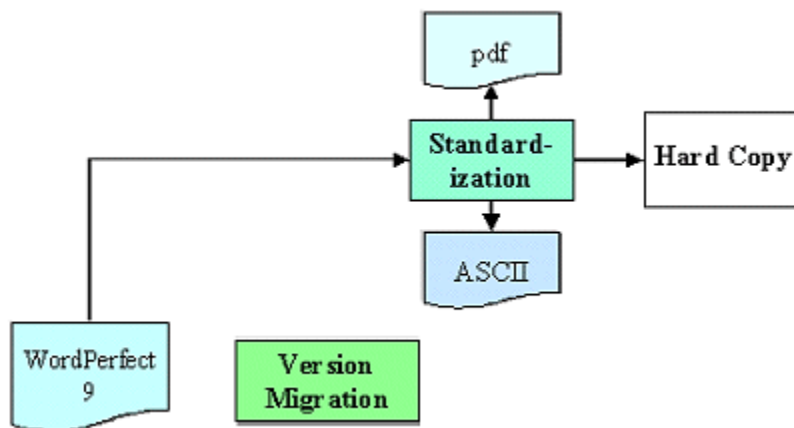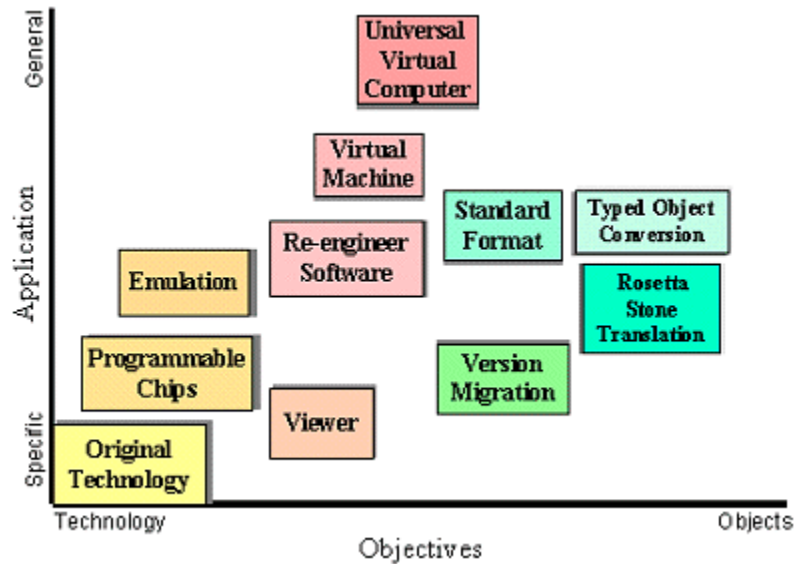
- Intellectual property rights
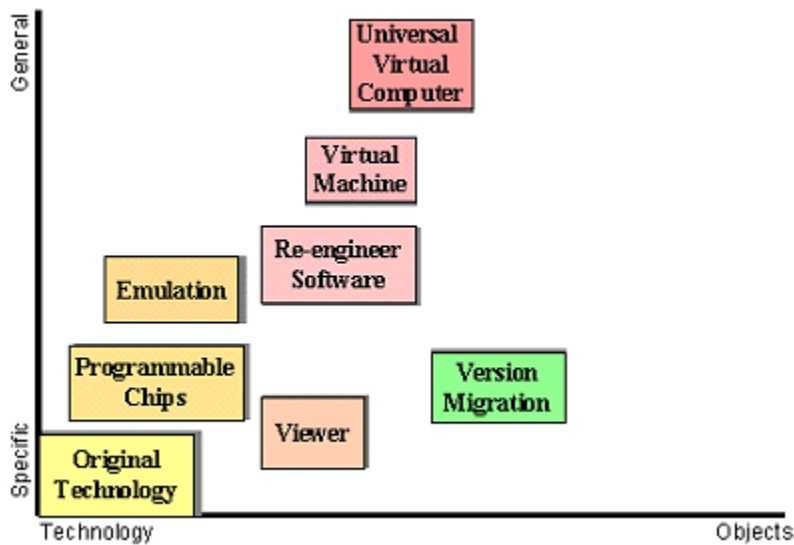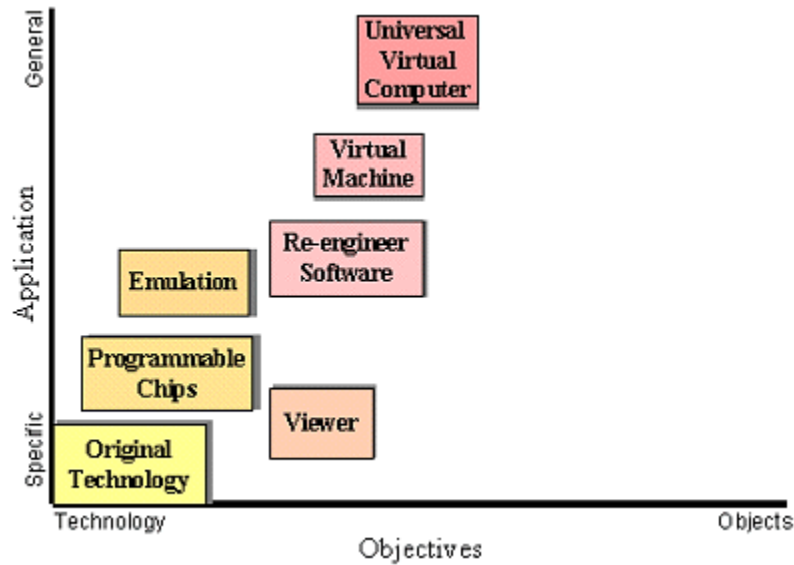
Version Migration

Migration to Standard Form

A two-axis chart. Vertical axis labeled "Application" with "General" at top and "Specific" at bottom. Horizontal axis labeled "Objectives" with "Technology" at left and "Objects" at right.

- Universal Virtual Computer
- Virtual Machine
- Standard Format
- Typed Object Conversion
- Re-engineer Software
- Emulation
- Rosetta Stone Translation
- Programmable Chips
- Version Migration
- Viewer
- Original Technology

Typed Object Model

### J. Ockerbloom, U. Pennsylvania

- All digital data 'things' are objects with
  - Specified attributes
  - Specified methods (operations)
  - Specific semantics
- A Type is defined by given values or parameters for the attributes, methods, or semantics of a class of objects.
- Any digital object (byte sequence) has a format; i.e. a specified encoding for an object of that type
- Byte sequences can be converted from one format to another
- It is possible to perform conversions that respect type
- 'Respectful' conversions cannot be distinguished when viewed through an interface for the type.

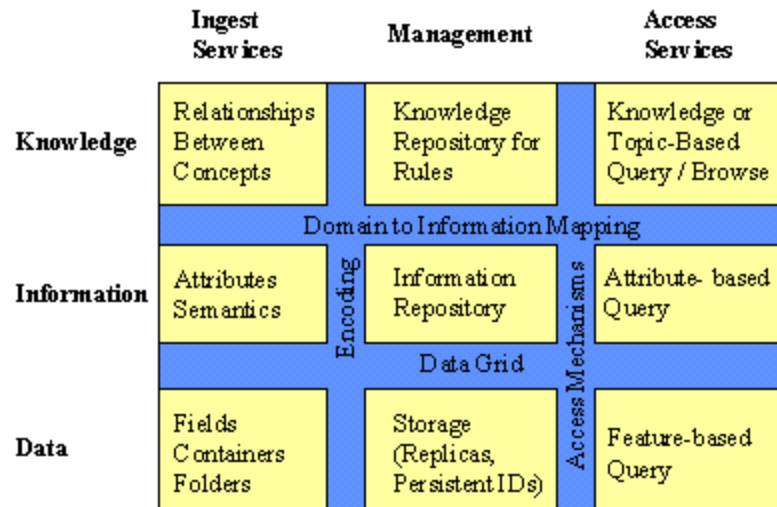## Rosetta Stone

Arcot Rajasekar, SDSC
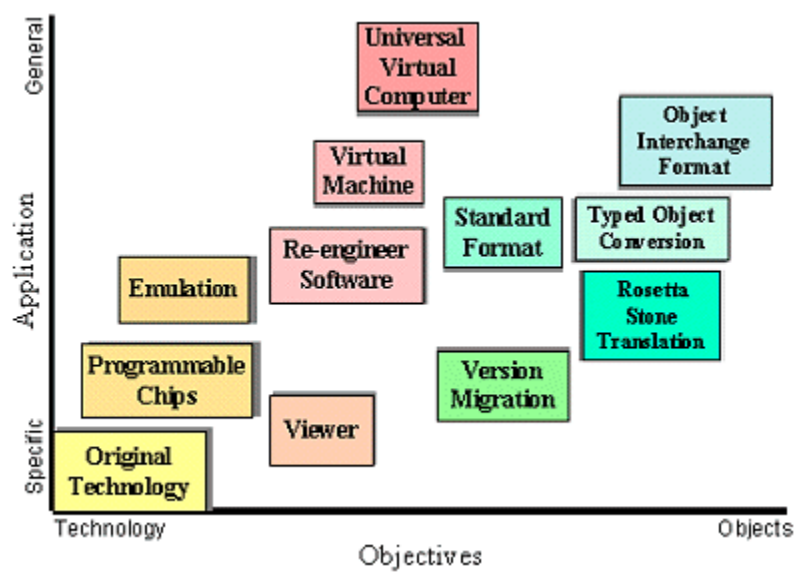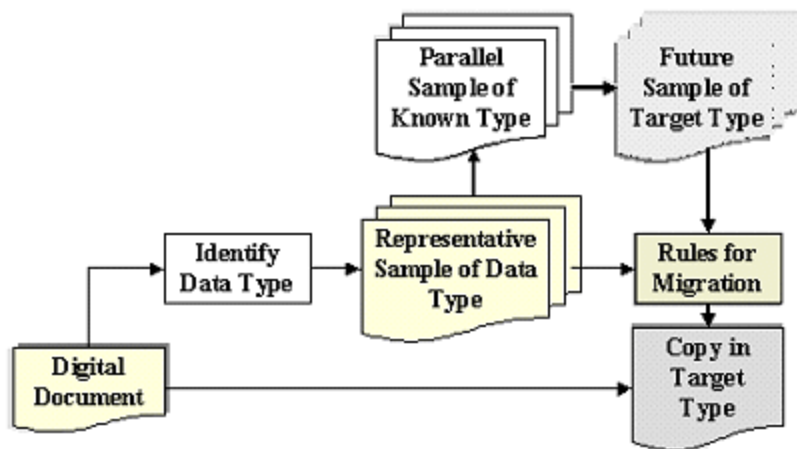
Object Interchange Format
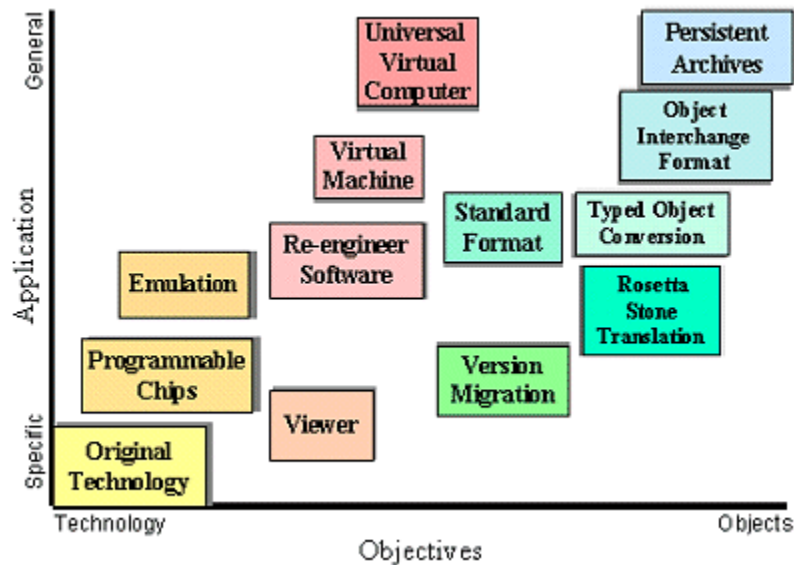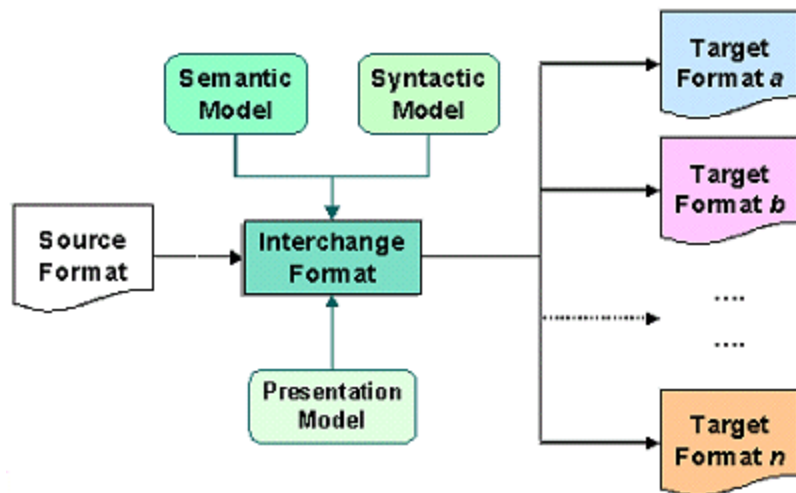
Persistent Archives Method

- Characterize significant properties of the things that are to be preserved.

- Express these properties in formal models

- Encapsulate objects in metadata defined in the models.

- Use software "mediators" to enable future technologies to interpret the models and metadata

  - to rebuild and repopulate collections
  - to re-present the records
  - support information discovery and delivery.

Knowledge Based Persistent Archives

	Ingest Services	Management	Access Services
Knowledge	Relationships Between Concepts	Knowledge Repository for Rules	Knowledge or Topic-Based Query / Browse
	Domain to Information Mapping		
Information	Attributes Semantics	Information Repository	Attribute- based Query
	Data Grid		
Data	Fields Containers Folders	Storage (Replicas, Persistent IDs)	Feature-based Query

Encoding ... *Access Mechanisms*

National Partnership for Advanced National Partnership for Advanced Computational Infrastructure