Building the Archives of the Future: Self-Describing Records

Kenneth Thibodeau

Director, Electronic Records Archives Program

National Archives and Records Administration

July 18, 2001

The Electronic Records Archives Vision

- Overcome technological obsolescence in a way that preserves demonstrably authentic records.
- Build a dynamic solution that incorporates the expectation of continuing change in information technology and in the records it produces.
- Find ways to take advantage of continuing progress in information technology in order to maintain and improve both performance and customer service

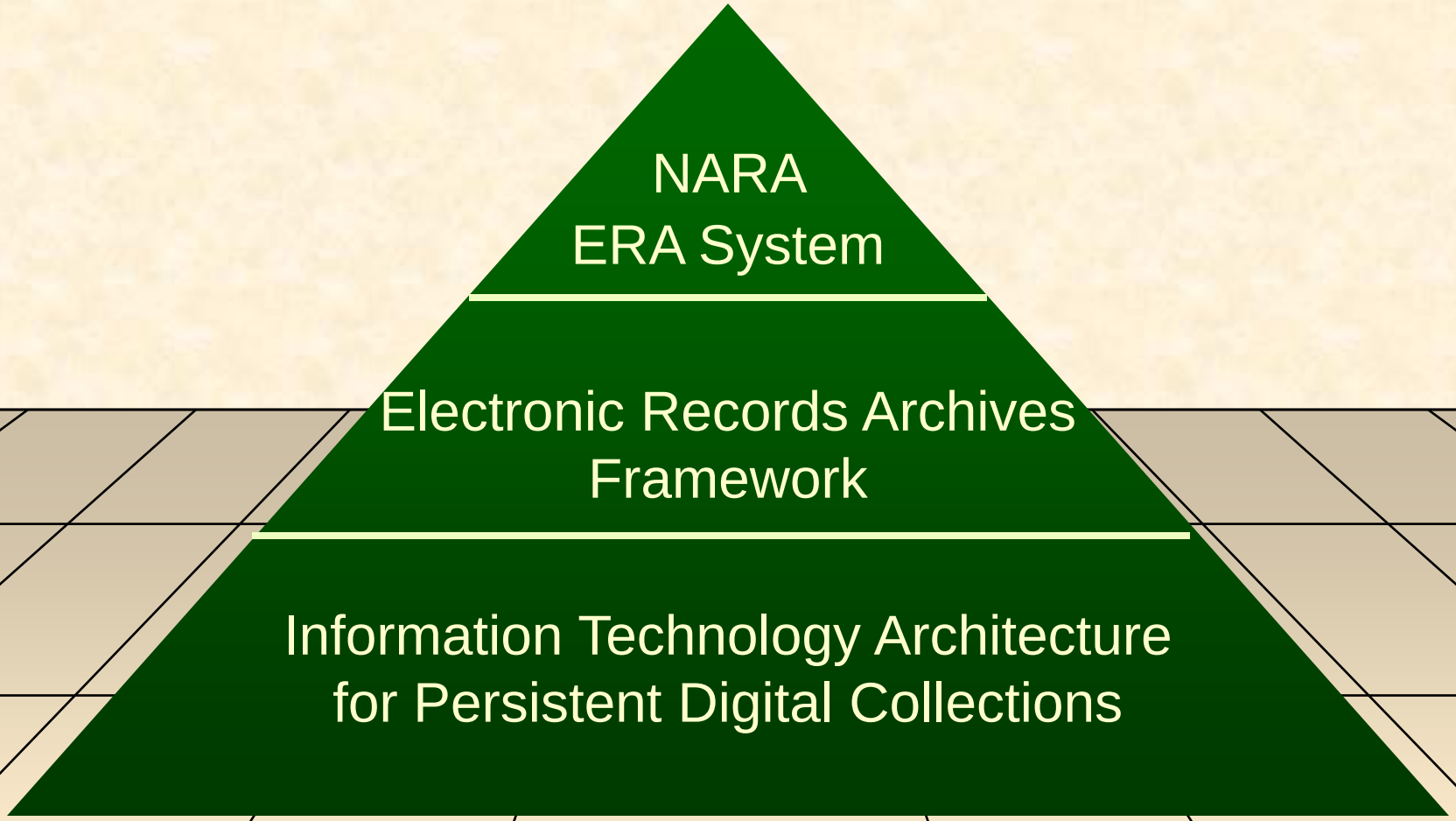
Electronic Records Archives Program

Critical Challenge

- Proven methods for preserving digital information across generations of technology are limited to the simplest formats
- Available methods are increasingly inadequate
- The market has not delivered solutions.

Electronic Records Archives Program

How will we develop the Electronic Records Archives?



ERA Infrastructure Concept

SCALABLE

Gb/sec Internet

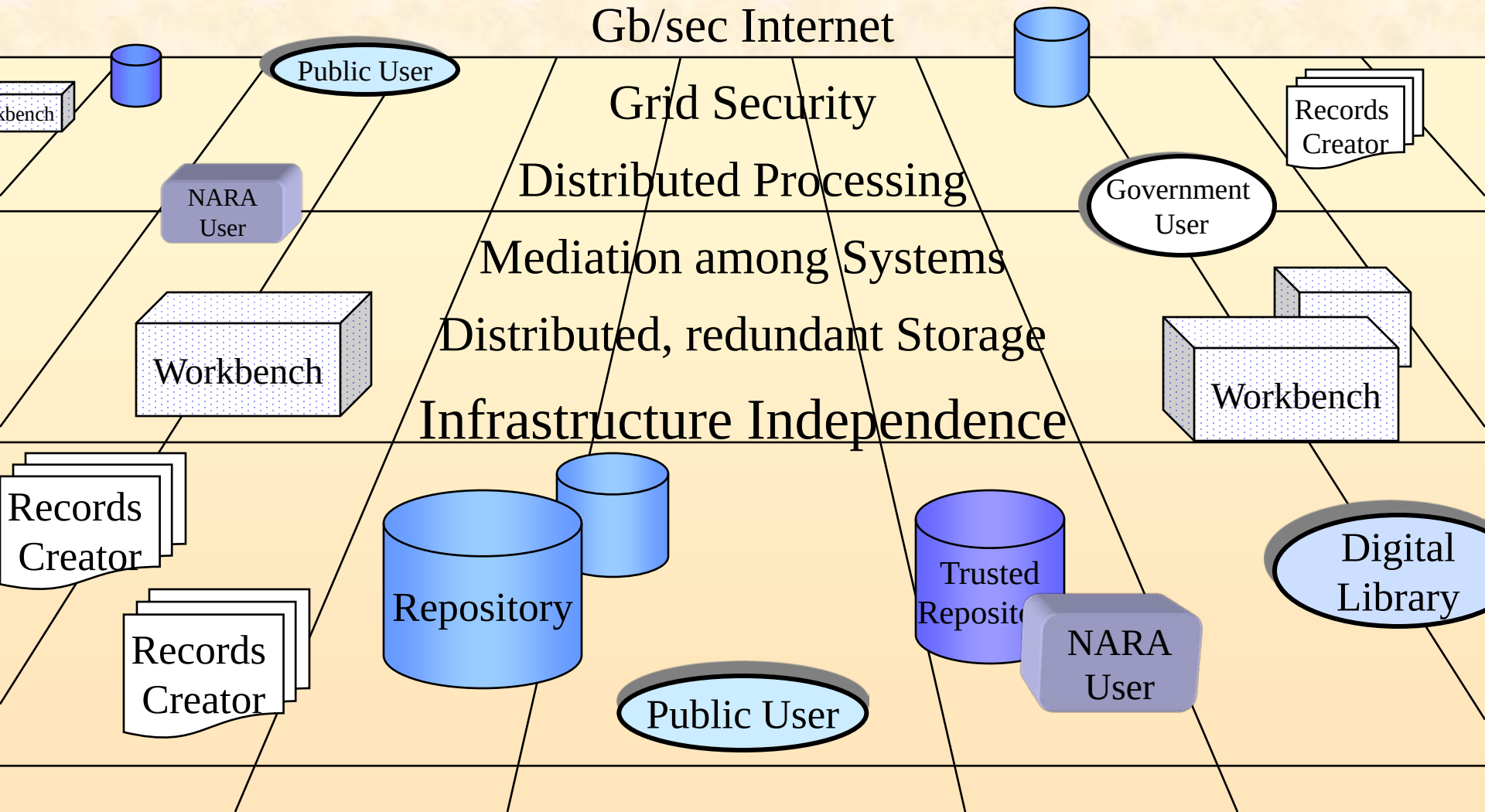
Grid Security

Distributed Processing

Mediation among Systems

Distributed, redundant Storage

Infrastructure Independence



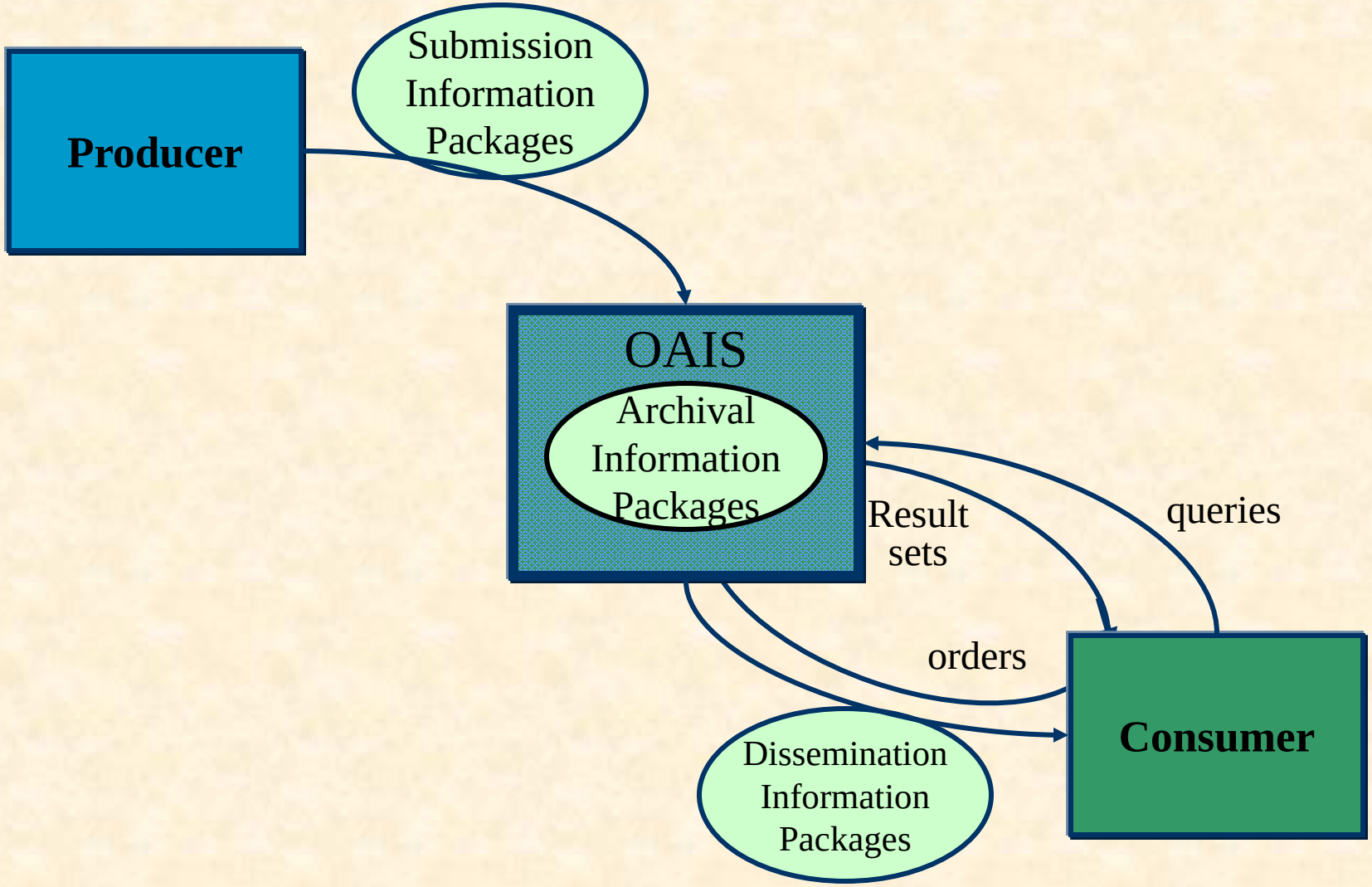
ERA Infrastructure

- In NARA (using NARANET)
  - **Archival workstations for staff**
  - **Reference workstations for researchers**
- On the National Information Infrastructure, under NARA's control
  - **ERA Ingest & Distribution portal (Internet & Media)**
  - **POP repositories (Normal, Trusted, Special)**
  - **Affiliated Archives**
- On the National Information Infrastructure
  - **Agency systems with access to NARA portal**
  - **Digital Libraries**
  - **Public Users**

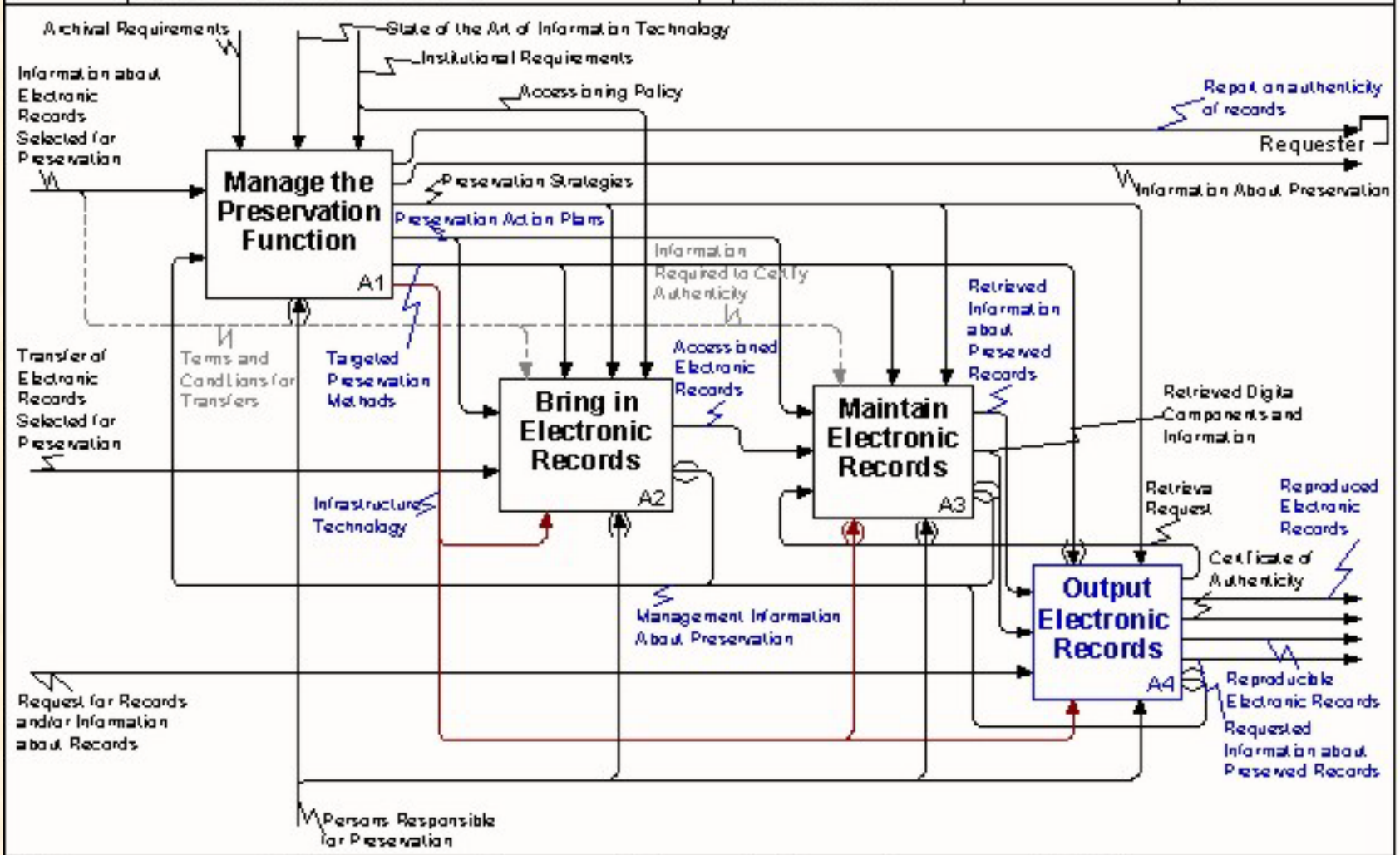
NARA Partnerships

- **Open Archival Information System** (OAIS) Reference Model
  - NASA, Consultative Committtee on Space Data Systems
- **Distributed Object Computation Testbed** (DOCT)
  - Defense Advanced Research Projects Agency, U.S. Patent and Trademark Office
- **National Partnership for Advanced Computational Infrastructure** (NPACI)
  - National Science Foundation
- **Presidential Electronic Records Processing Operational System** (PERPOS)
  - Army Research Laboratory, Georgia Tech Research Institute
- **Archivist's Workbench**
  - NHPRC Grant to San Diego Supercomputer Center
- **International research on Permanent Authentic Records in Electronic Systems** (InterPARES)
  - 7 international, multidisciplinary research teams, 10 national archives

ERA Functional Model
An Open Archival Information System Implementation



USED AT: Workshop 8	AUTHOR: Preservation Task Force	DATE: 2/17/2000	WORKING	READER DATE	CONTEXT:
	PROJECT: InterPARES Project	REV: 7/6/2001	DRAFT		
			RECOMMENDED		
	NOTES: 1 2 3 4 5 6 7 8 9 10		PUBLICATION		A-0

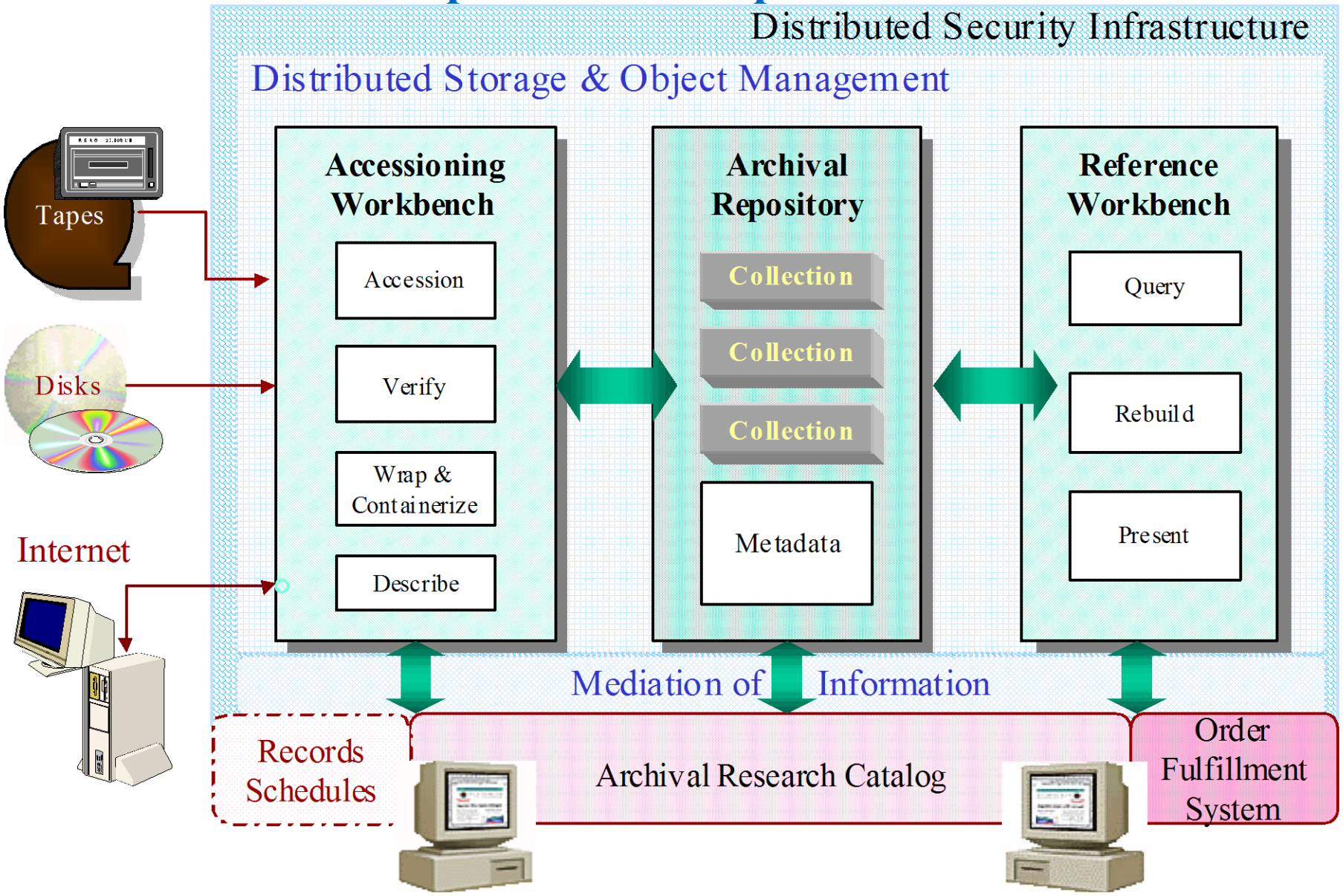


NODE:	TITLE:	NUMBER:
A0	Preserve Electronic Records	v 5b draft

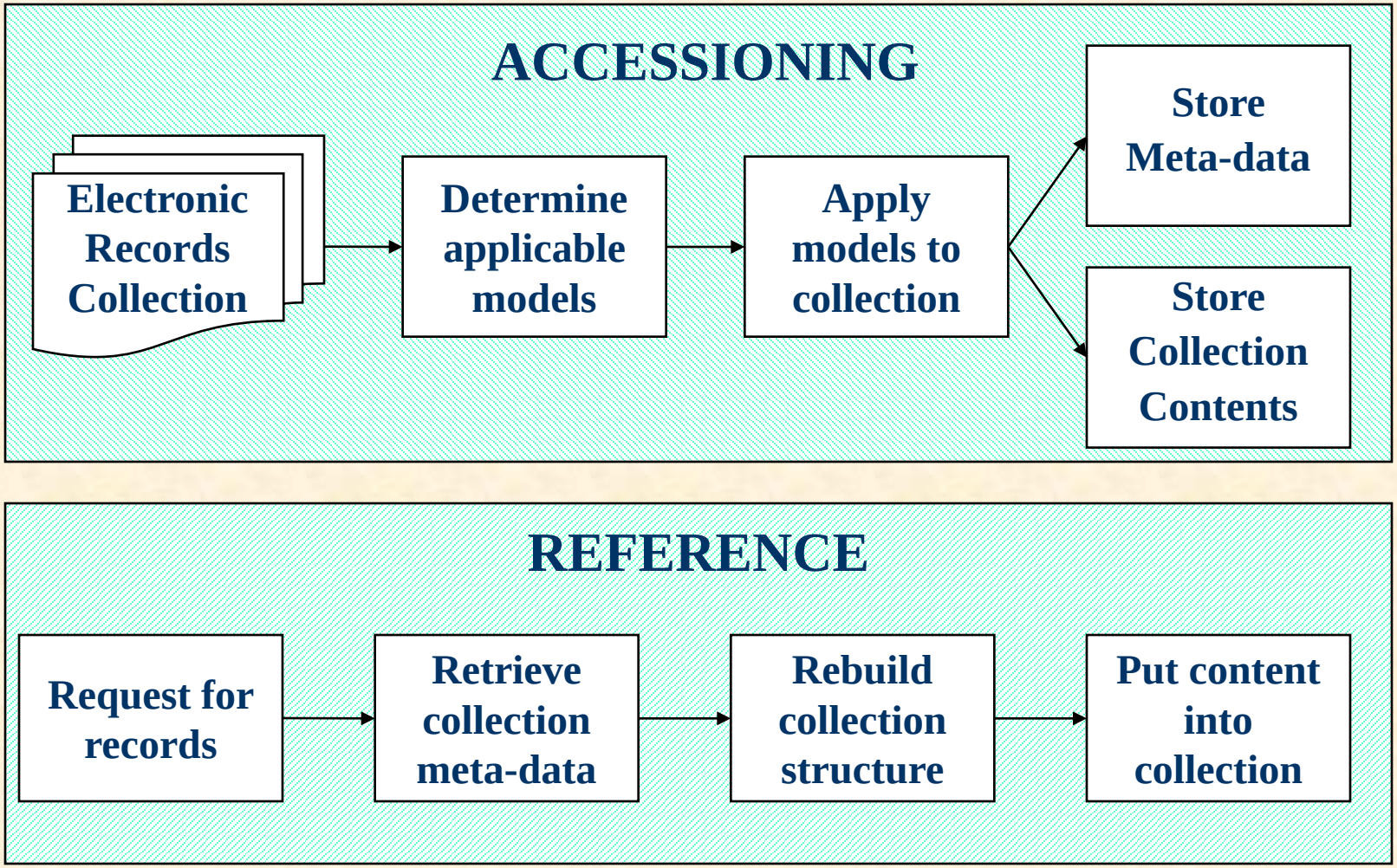
Information Management Architecture for Persistent Object Preservation

	Ingest Services		Management		Access Services
Knowledge	Relationships Between Concepts	XTM DTD	Knowledge Repository for Rules	Rules - KQL	Knowledge or Topic-Based Query / Browse
	(Topic Maps / Buckets / Model-based Access)				
Information	Attributes Semantics	XML DTD	Information Repository	SDLIP	Attribute- based Query
	(Data Handling System - SRB / FTP / HTTP)				
Data	Fields Containers Folders	MCAT/HDF	Storage (Replicas, Persistent IDs)	Grids	Feature-based Query

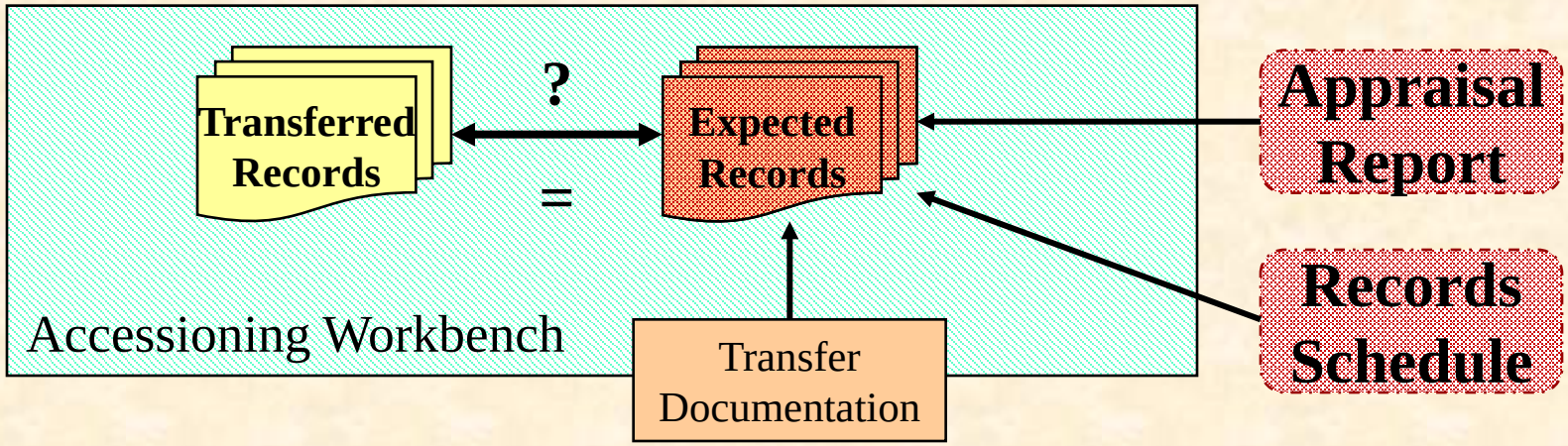
ERA: Archival Components Concept



ERA Processes



Accept an Accession?



- What should the agency have transferred?
- What did the agency say it transferred?
- What was transferred?

SELF-DESCRIBING

- Records
- Files
- Series
- Record Systems

Persistent Object 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.

E-mail: Groupwise view

Mail From: Yigal Arens <arens@ISI.EDU>

File Edit View Actions Tools Window Help

From: Yigal Arens <arens@ISI.EDU>

CC:

To:

BC: Ken Thibodeau

Subject: Announcing DG Online, the magazine of digital governmentresearch

Message:

DG Online <http://www.dgrc.org/dg-online/>.

DG Online: The Magazine for Digital Government Research is the new online quarterly of dg.o (DigitalGovernment.Org), a national consortium of government agencies, computer science researchers, the IT industry, and civic organizations concerned with improving online government operations and services.

DG Online presents the latest developments in advanced computer and IT research for Digital Government along with news and viewpoints on the most important DG issues:

* Cooperation among federal, state, and local government agencies
* Privacy and security
* Universal access--bridging the digital divide
* Streamlined delivery and tracking of public services
* Electronic voting, taxes, the Census, and other sensitive online data collection
* User-friendly interfaces

Attach:

Close

Reply

Forward

Info

Delete

Date: 3/28/01 11:07PM

E-mail: text editor view

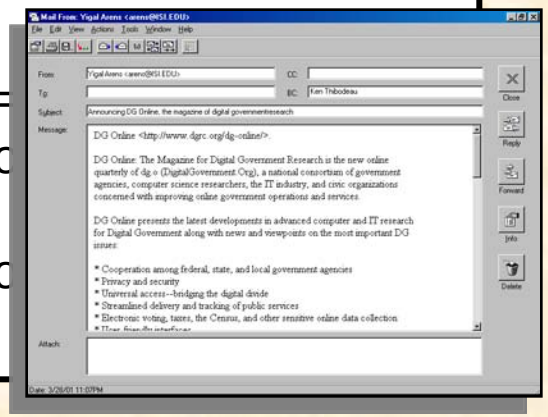
```
ÿWPCL    ■■ ■     Ûÿ■ 2      ■ ■    B    ■ ■    J                          ■#| x      Ð■■      °■°■■ ■ÐÐ■■
 ■ ■ÐÐ■Ð
             0    p■ÿÿÿ
```

E-mail: MIME-aware view

```
Message-Id: <p05010429b6e7e938e0b4@[10.2.68.205]>
X-Organization: USC/Information Sciences Institute
X-Phone: (310) 822-1511  ext. 766
X-Fax: (310) 822-0751
Date: Wed, 28 Mar 2001 22:22:30 -0500
To: Readers:;
From: Yigal Arens <arens@ISI.EDU>
Subject: Announcing DG Online, the magazine of
digital government
 research
Content-Type: multipart/alternative;
boundary="============_-
1226283781==_ma============"

--============_-1226283781==_ma============
Content-Type: text/plain; charset="us-ascii" ;
format="flowed"

DG Online <http://www.dgrc.org/dg-online/>.

DG Online: The Magazine for Digital Government
Research is the new online
```

Tagged MIME E-mail Message

```
<Message-Id> p05010429b6e7e938e0b4@[10.2.68.205]
</Message-Id>
<X-Organization>USC/Information Sciences Institute</X-
Organization>
<X-Phone>(310) 822-1511  ext. 766</X-Phone>
<X-Fax>(310) 822-0751</X-Fax>
<Date>Wed, 28 Mar 2001 22:22:30 -0500</Date>
<To>Readers:;</To>
<From>Yigal Arens {arens@ISI.EDU}</From>
<Subject>Announcing DG Online, the magazine of digital
government research</Subject>
<Content-Type: multipart/alternative;
boundary="============_-1226283781==_ma============"

--============_-1226283781==_ma==========
<Content-Type: text/plain; charset="us-asc
format="flowed">
<Message_Body>DG Online {http://www.dgrc.c
....
```



Structure of E-mail Message
aka: Document Type Definition

```
<!ELEMENT Email_Message (Header, Message Body, Attachment*)>
    <!ELEMENT Header (Internal Header, External Header)>
        <!ELEMENT Internal Header     (Message_Id, X-
          Organization, X-Phone, X-Fax)>
            <!ELEMENT Message-Id>
            <!ELEMENT X-Organization>
            <!ELEMENT X-Phone>
            <!ELEMENT X-Fax>
        <!ELEMENT External Header     (Date, To, From,
          Subject)>
            <!ELEMENT Date     (Weekday, Day_of_Month, Month,
              Time>
            <!ELEMENT To                (#PCDATA)+>
            <!ELEMENT From     (#PCDATA)>
            <!ELEMENT Subject  (#PCDATA)*>
    <!ELEMENT Message_Body     (#PCDATA)*>
```

File of Electronic Records

Types of Records

REPORT OF INDEPENDENT ACCOUNTANTS

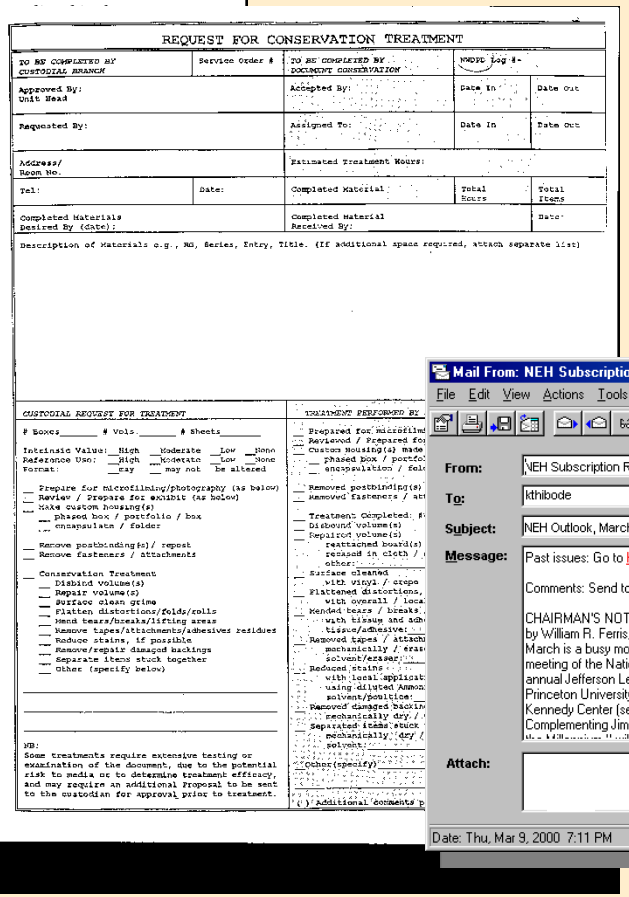
To the Board of Directors and Stockholders of Great Plains Software, Inc.

In our opinion, the consolidated financial stat
accompanying index present fairly, in all materi
of Great Plains Software, Inc. and its subsidiari
and the results of their operations and their cash
years in the period ended May 31, 1999, in conf
accepted accounting principles. In addition, in c
statement schedules listed in the accompanying
material respects, the information set forth there
with the related consolidated financial statemen
and financial statement schedules are the respor
management; our responsibility is to express an
statements and financial statement schedules ba
conducted our audits of these statements in acc
accepted auditing standards, which require that
to obtain reasonable assurance about whether th
of material misstatement. An audit includes exa
supporting the amounts and disclosures in the fi
the accounting principles used and significant e
and evaluating the overall financial statement p
audits provide a reasonable basis for the opinio
/s/ PricewaterhouseCoopers LLP

PricewaterhouseCoopers LLP

Minneapolis , Minnesota

June 25, 1999



REQUEST FOR CONSERVATION TREATMENT

TO BE COMPLETED BY CUSTODIAL BRANCH	Service Order #	TO BE COMPLETED BY DOCUMENT CONSERVATION	NNDPC Log #	
Approved By: Unit Head		Accepted By:	Date In	Date Out
Requested By:		Assigned To:	Date In	Date Out
Address/ Room No.		Estimated Treatment Hours:		
Tel:	Date:	Completed Material	Total Hours	Total Items
Completed Materials Desired By (date):		Completed Material Received By:		Date:

Description of Materials e.g., RG, Series, Entry, Title. (If additional space required, attach separate list)

CUSTODIAL REQUEST FOR TREATMENT

Boxes_____ # Vols._____ # Sheets_____

Intrinsic Value: __High __Moderate __Low __None
Reference Use: __High __Moderate __Low __None
Format: __ray __ may not be altered

__ Prepare for microfilming/photography (as below)
__ Review / Prepare for exhibit (as below)
__ Make custom housing(s)
__ phased box / portfolio / box
__ encapsulate / folder

__ Remove postbinding(s) / repost
__ Remove fasteners / attachments

__ Conservation Treatment
__ Disbind volume(s)
__ Repair volume(s)
__ Surface clean grime
__ Flatten distortions/folds/rolls
__ Mend tears/breaks/lifting areas
__ Remove tapes/attachments/adhesives residues
__ Reduce stains, if possible
__ Remove/repair damaged backings
__ Separate items stuck together
__ Other (specify below)

NB:
Some treatments require extensive testing or examination of the document, due to the potential risk to media or to determine treatment efficacy, and may require an additional Proposal to be sent to the custodian for approval prior to treatment.

TREATMENT PERFORMED BY

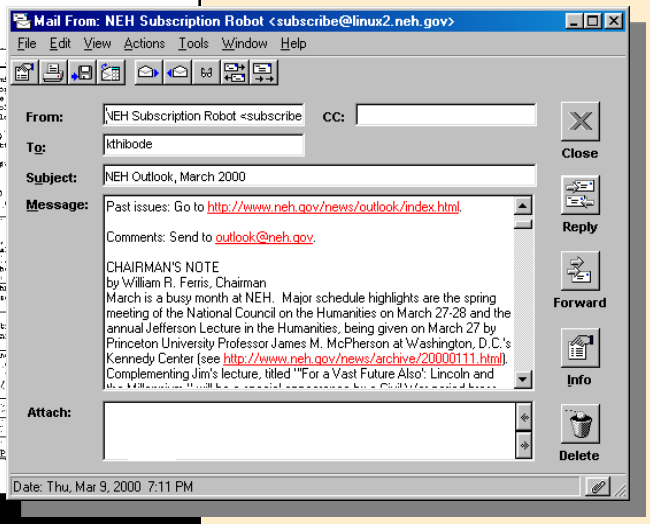
__ Prepared for microfilm
__ Reviewed / Prepared fo
__ Custom housing(s) made
__ phased box / portfo
__ encapsulation / fol

__ Removed postbinding(s)
__ Removed fasteners / at

__ Treatment Completed: #
__ Disbound volume(s)
__ Repaired volume(s)
reattached board(s)
reglued in cloth /
other:

__ Surface cleaned
with vinyl / crepe
__ Flattened distortions,
with overall / loca
__ Mended tears / breaks,
with tissue and adh
tissue/adhesive:
__ Removed tapes / attach
mechanically / eras
solvent/eraser
__ Reduced stains
with local applicat
using diluted Ammon
solvent/poultice:
__ Removed damaged backin
mechanically dry /
__ Separated items stuck
mechanically dry /
solvent:
__ Other (specify)

() Additional comments p



Mail From: NEH Subscription Robot <subscribe@linux2.neh.gov>

File Edit View Actions Tools Window Help

From: NEH Subscription Robot <subscribe
CC:
To: kthibode
Subject: NEH Outlook, March 2000
Message: Past issues: Go to http://www.neh.gov/news/outlook/index.html.

Comments: Send to outlook@neh.gov.

CHAIRMAN'S NOTE
by William R. Ferris, Chairman
March is a busy month at NEH. Major schedule highlights are the spring meeting of the National Council on the Humanities on March 27-28 and the annual Jefferson Lecture in the Humanities, being given on March 27 by Princeton University Professor James M. McPherson at Washington, D.C.'s Kennedy Center (see http://www.neh.gov/news/archive/20000111.html). Complementing Jim's lecture, titled "'For a Vast Future Also': Lincoln and

Attach:

Close
Reply
Forward
Info
Delete

Date: Thu, Mar 9, 2000 7:11 PM

eXtensible Business Reporting Language (XBRL) Example

REPORT OF INDEPENDENT ACCOUNTANTS

To the Board of Directors and Stockholders of Great Plains Software, Inc.

In our opinion, the consolidated financial statements listed in the accompanying index present fairly, in all material respects, the financial position of Great Plains Software, Inc. and its subsidiaries at May 31, 1999 and 1998, and the results of their operations and their cash flows for each of the three years in the period ended May 31, 1999, in conformity with generally accepted accounting principles. In addition, in our opinion, the financial statement schedules listed in the accompanying index present fairly, in all material respects, the information set forth therein when read in conjunction with the related consolidated financial statements. These financial statements and financial statement schedules are the responsibility of the Company's management; our responsibility is to express an opinion on these financial statements and financial statement schedules based on our audits. We conducted our audits of these statements in accordance with generally accepted auditing standards, which require that we plan and perform the audit to obtain reasonable assurance about whether the financial statements are free of material misstatement. An audit includes examining, on a test basis, evidence supporting the amounts and disclosures in the financial statements, assessing the accounting principles used and significant estimates made by management, and evaluating the overall financial statement presentation. We believe that our audits provide a reasonable basis for the opinion expressed above.

/s/ PricewaterhouseCoopers LLP

GreatPlainsInstance - XML Notepad

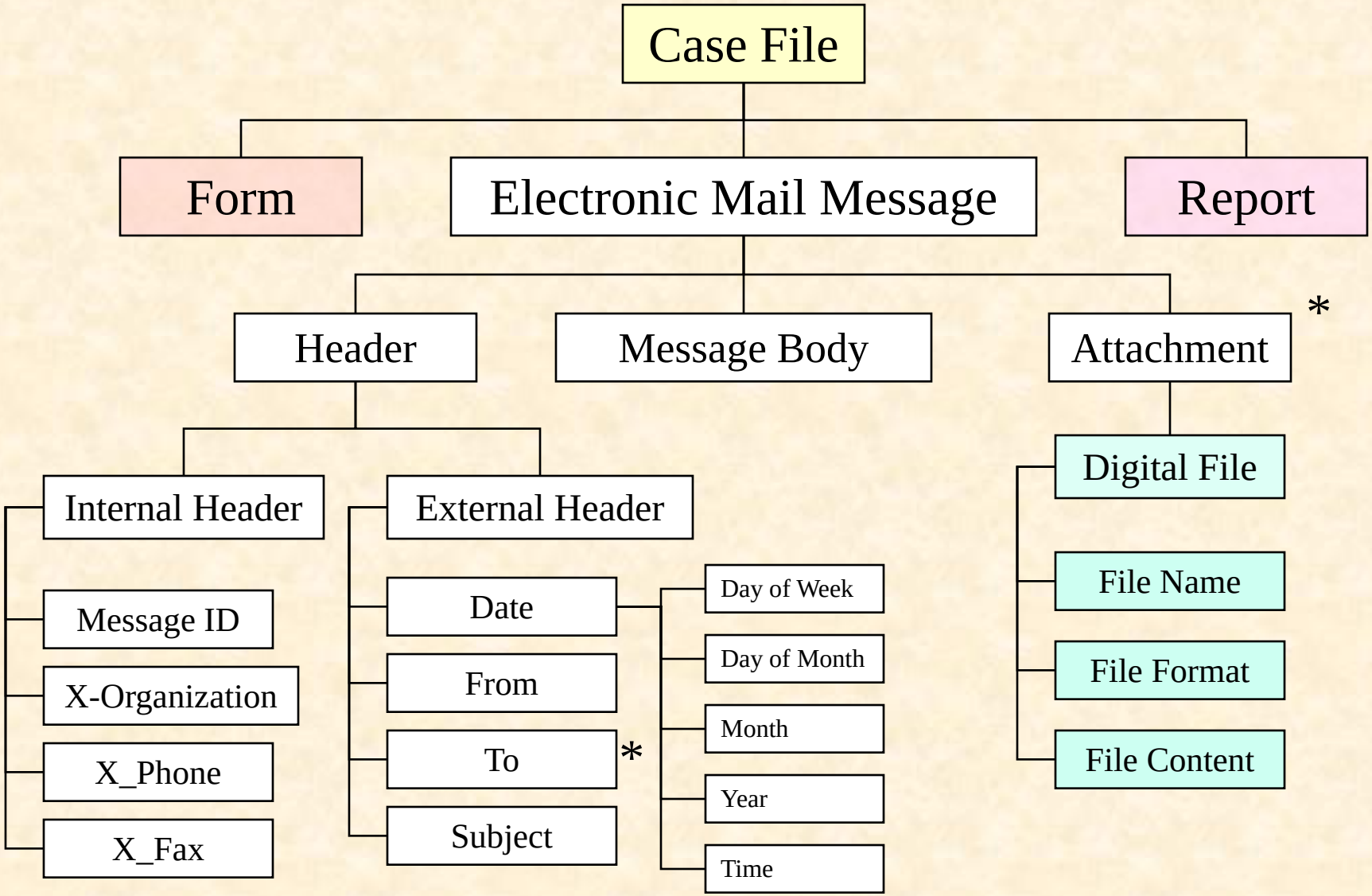
File Edit View Insert Tools Help

XBRL DTD

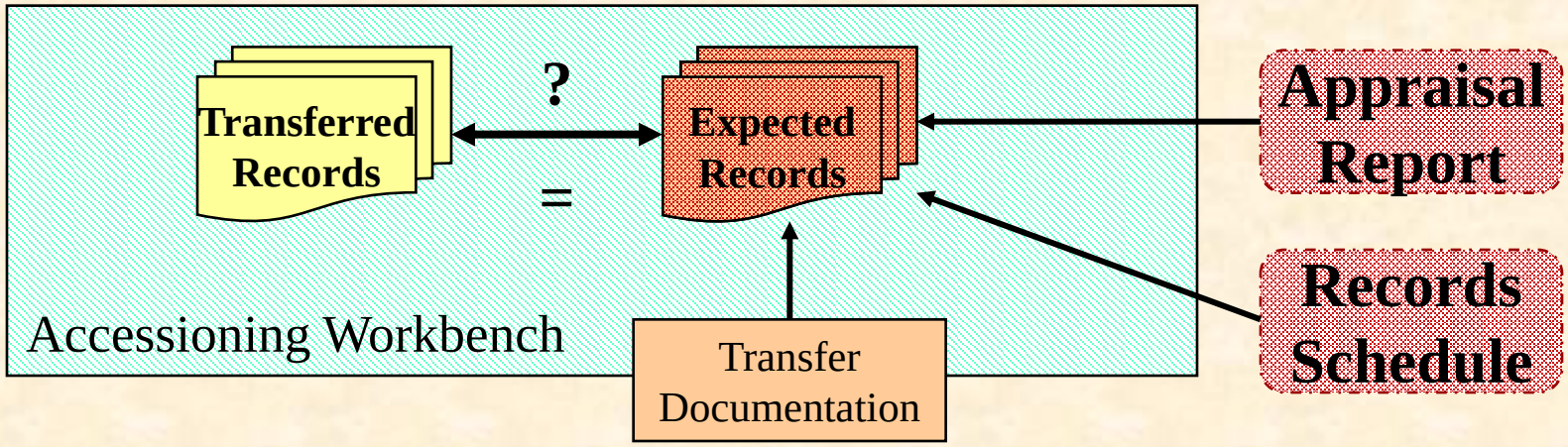
Structure	Values
[COMMENT]	SECTION: AccountantReport
ITEM	PricewaterhouseCoopers LLP
TYPE	accountant_sReport.independent
ITEM	/s/ PricewaterhouseCoopers LLP
TYPE	accountant_sReport.accountantSignature
ITEM	Minneapolis
TYPE	accountantSignature.city
ITEM	Minnesota
TYPE	accountantSignature.state
ITEM	REPORT OF INDEPENDENT ACCOUNTANTS
TYPE	accountant_sReport.titleOfAccountantsReport
ITEM	To the Board of Directors and Stockholders of Great Plains Software, Inc.
TYPE	accountant_sReport.addressee
ITEM	June 25, 1999
TYPE	reportDate.date
ITEM	Unqualified
TYPE	accountant_sReport.typeOfOpinion
ITEM	US GAAP
TYPE	reportingMethod.generallyAcceptedAccountingPrinciples
ITEM	In our opinion, the consolidated financial statements listed in the accompanying index present fairly, in all material respects, the f...
TYPE	scopeOfWorkPerformed.auditedFinancialStatements
[COMMENT]	SECTION: BalanceSheet
ITEM	26983
ID	BS-01
TYPE	cashAndCashEquivalents.cashEquivalents
PERIOD	1999-05-31
ITEM	18197
ITEM	96700
ITEM	48721
ITEM	12593
	8790

For Help, press F1 NUM

Structure expressed as Tree



Accept an Accession?

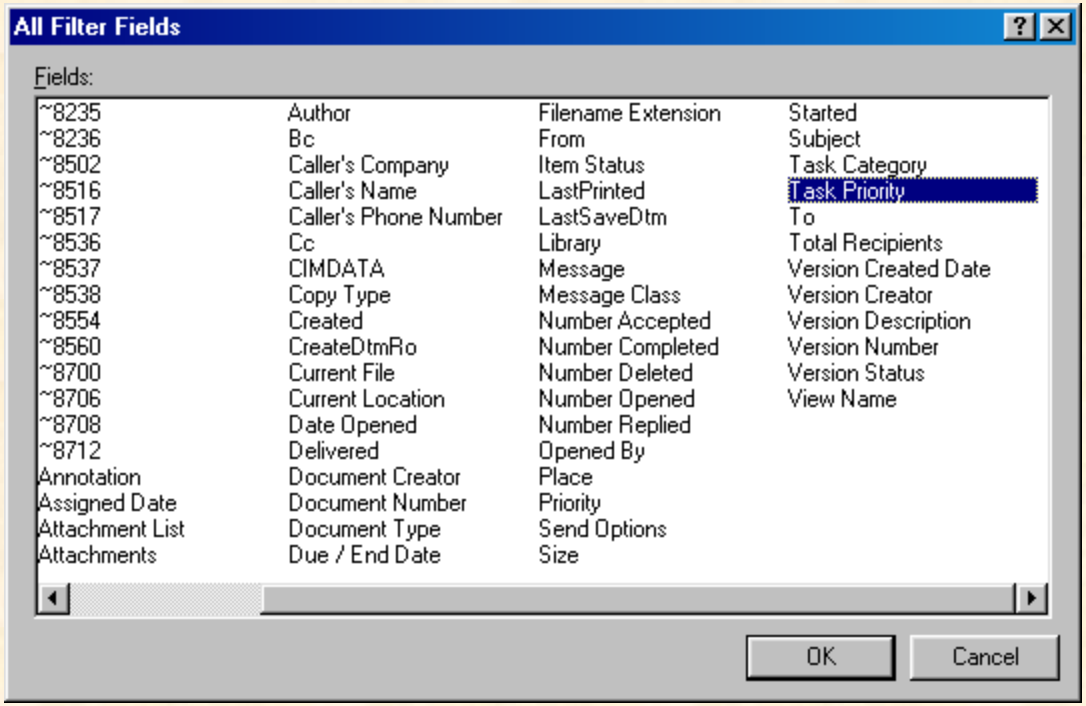


How does ERA determine the dates of records?

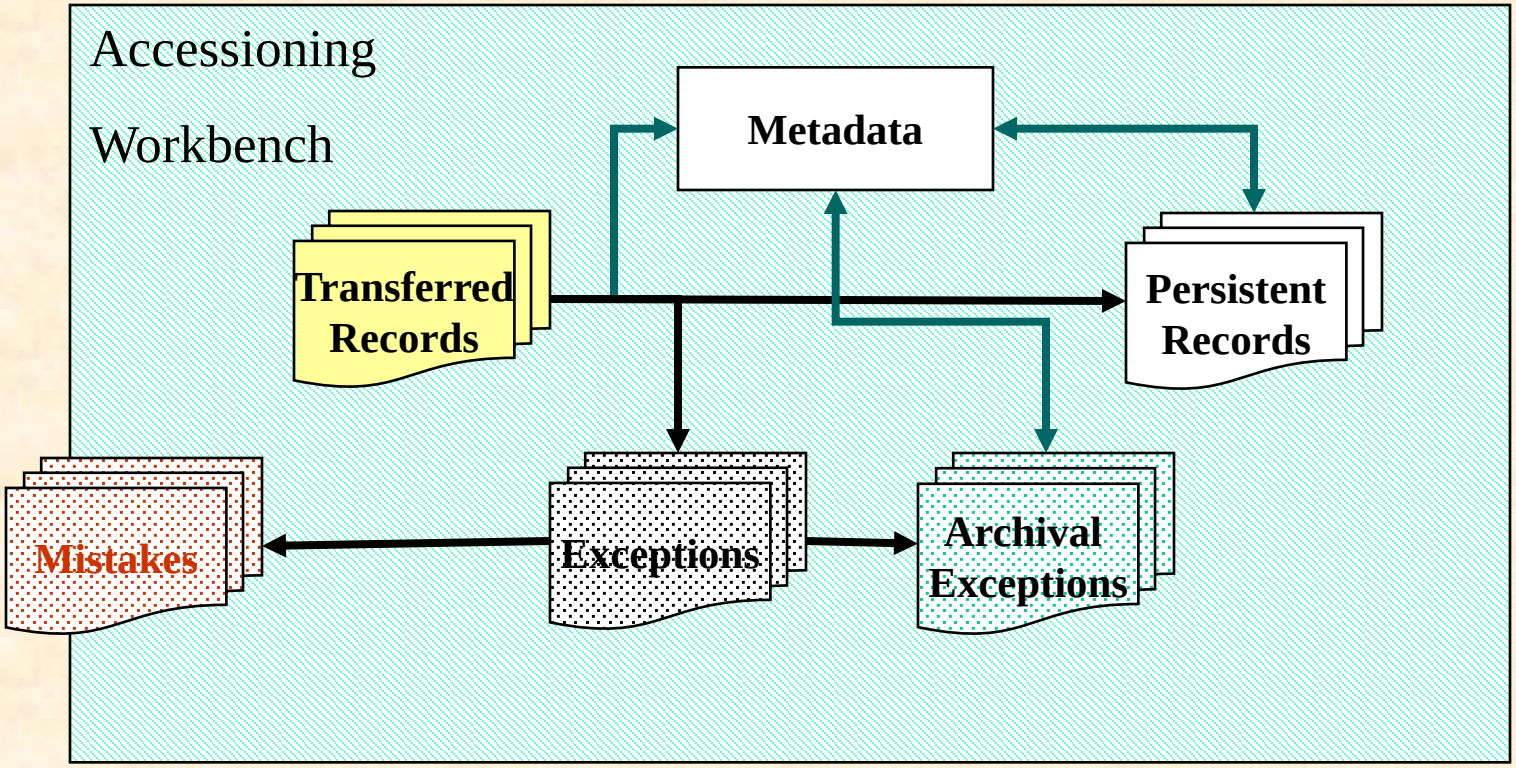
- E-mail
  - All e-mail contains a field indicating the date it was sent. For the sender, that is the date of the record. ERA needs to search the date-sent fields.
    - *(Technology solution)*
- Attachments to e-mail messages
  - Attachments to a record are parts of that record. The date of the message is the date of the record.
    - *(Archival principle)*
- Records forwarded, via e-mail, for filing in a recordkeeping system
  - E-mail is used only to transmit a record to the system. The date of the attached record depends on the record
    - *(Archival principle)*

Defining models for electronic records e.g. E-mail

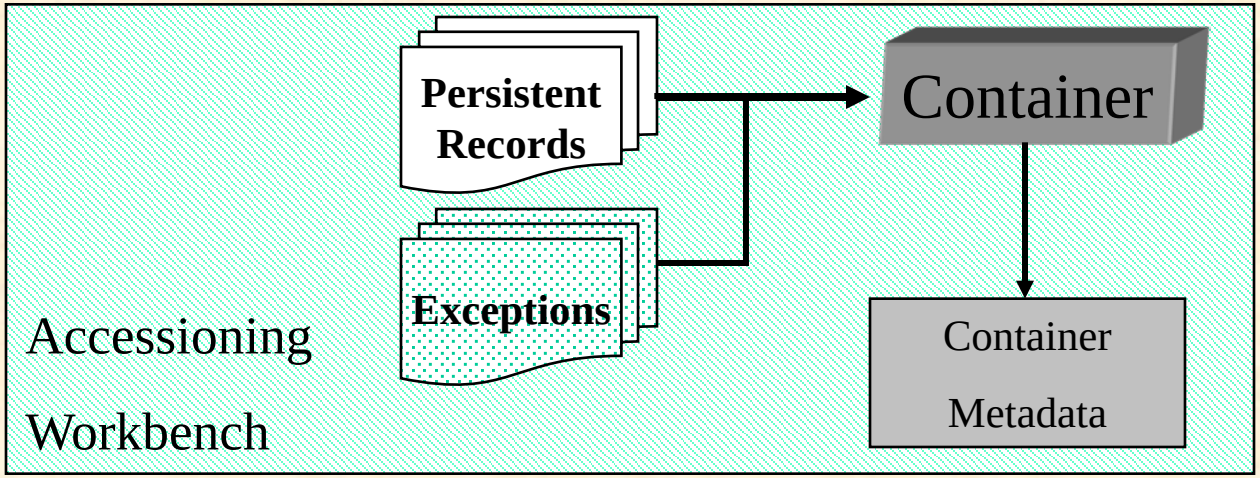
- All E-mail
  - Groupwise mail
  - cc:mail
  - USENET mail
    - User defined fields
  - …..



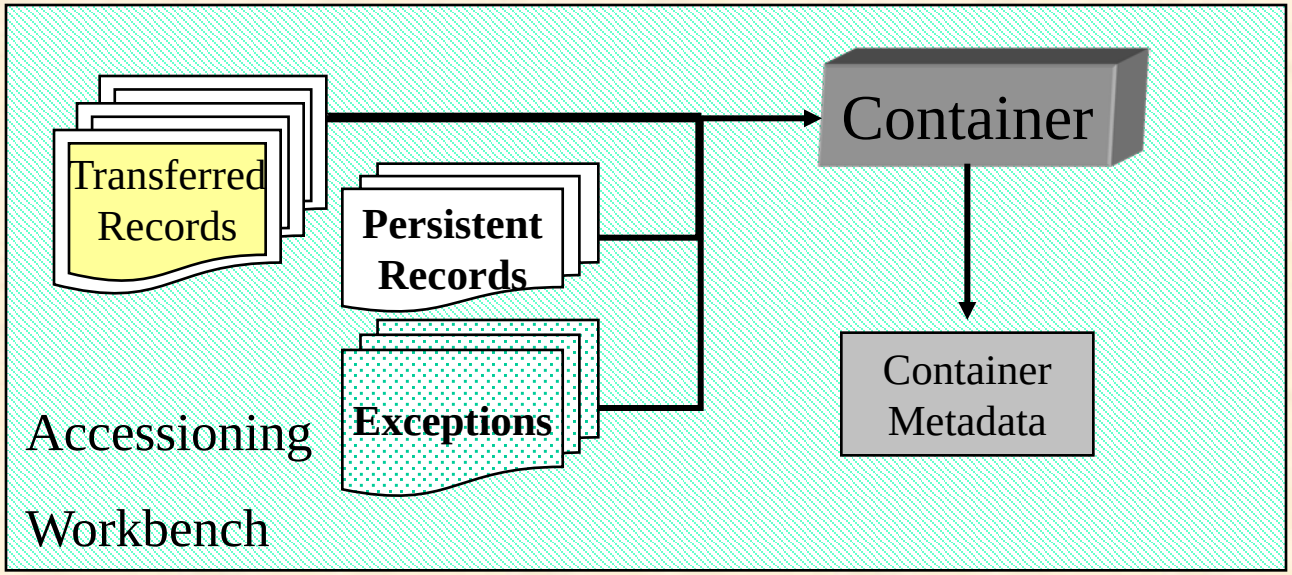
Transformation



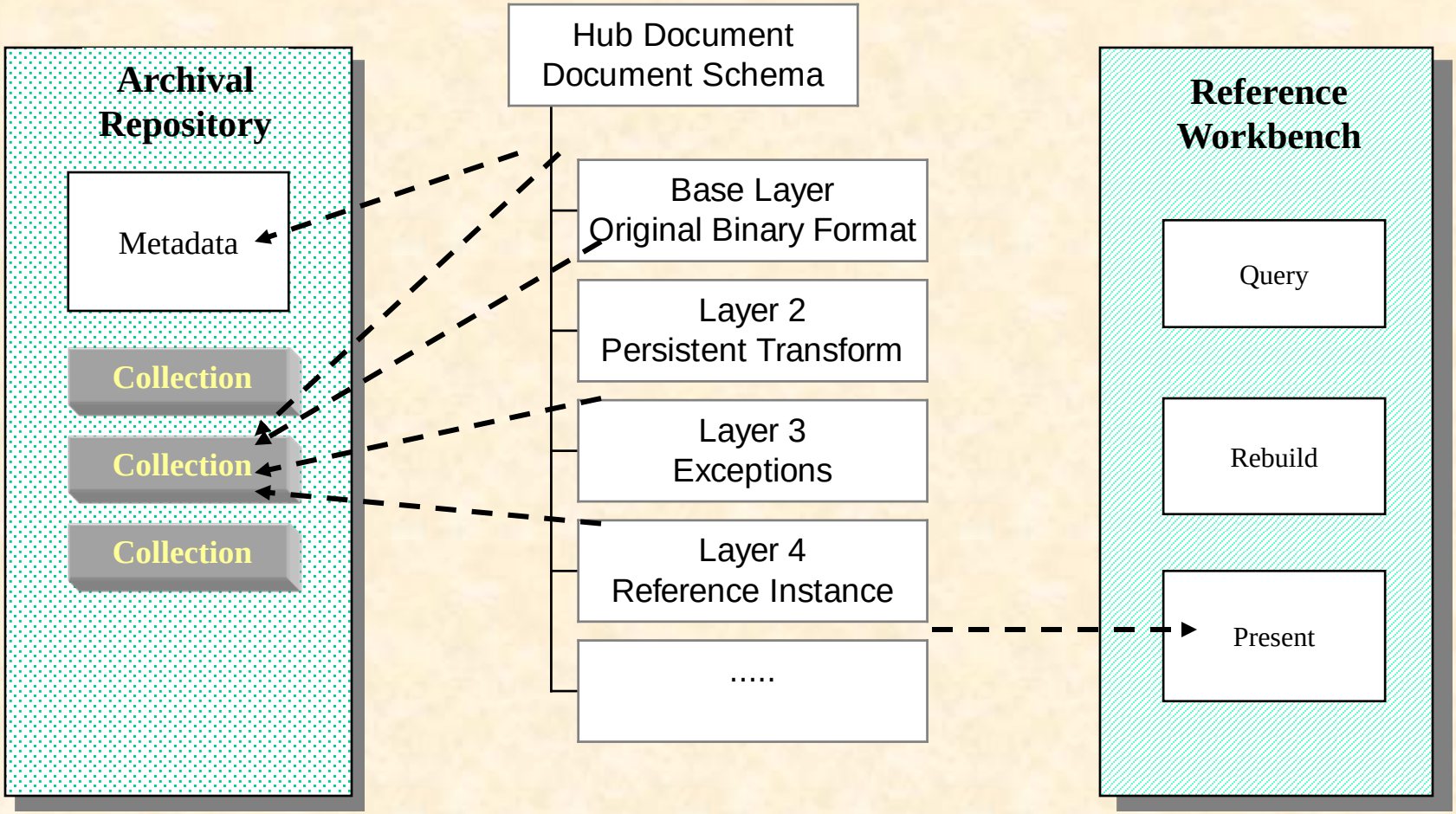
Aggregation



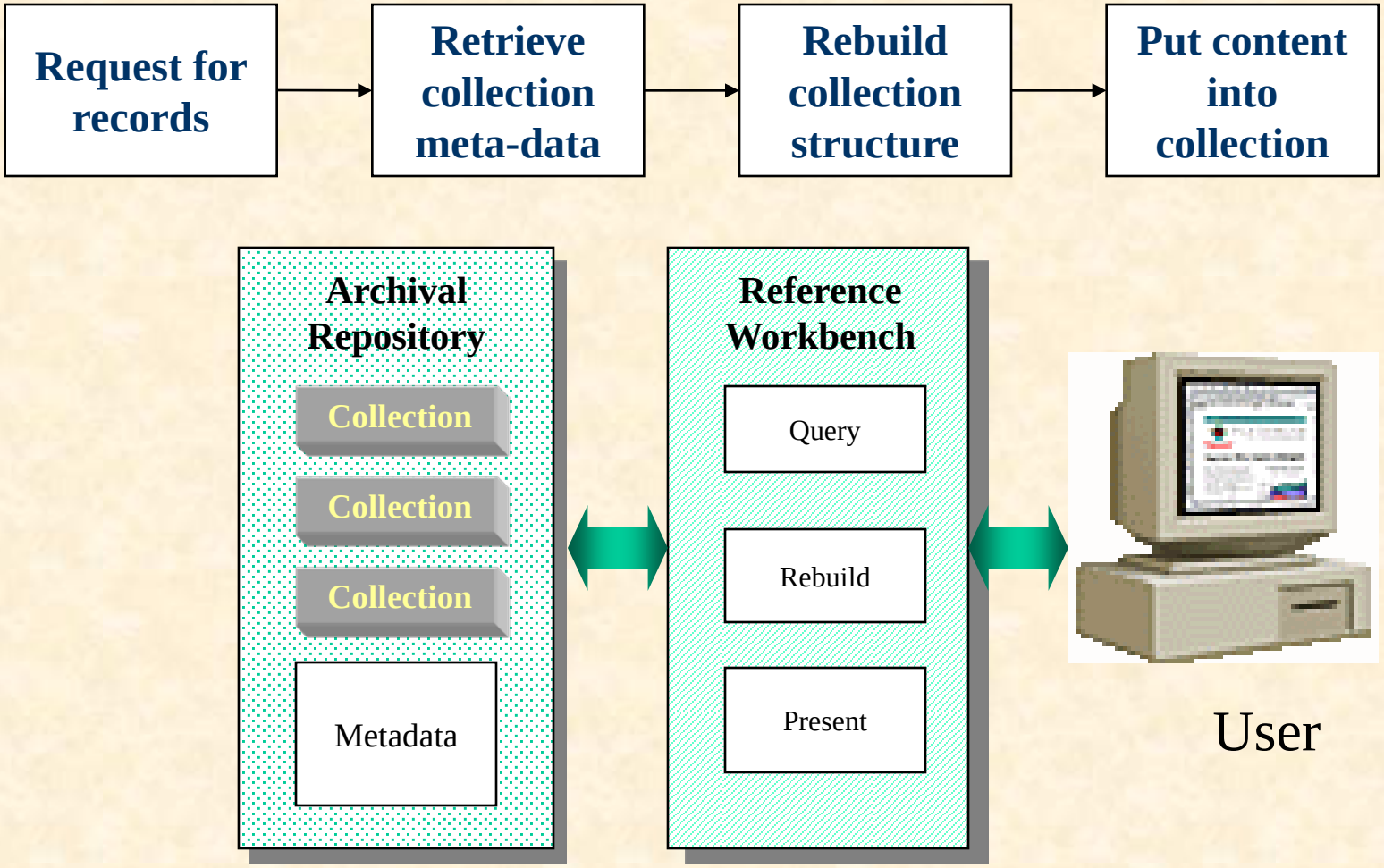
Aggregation: risk management



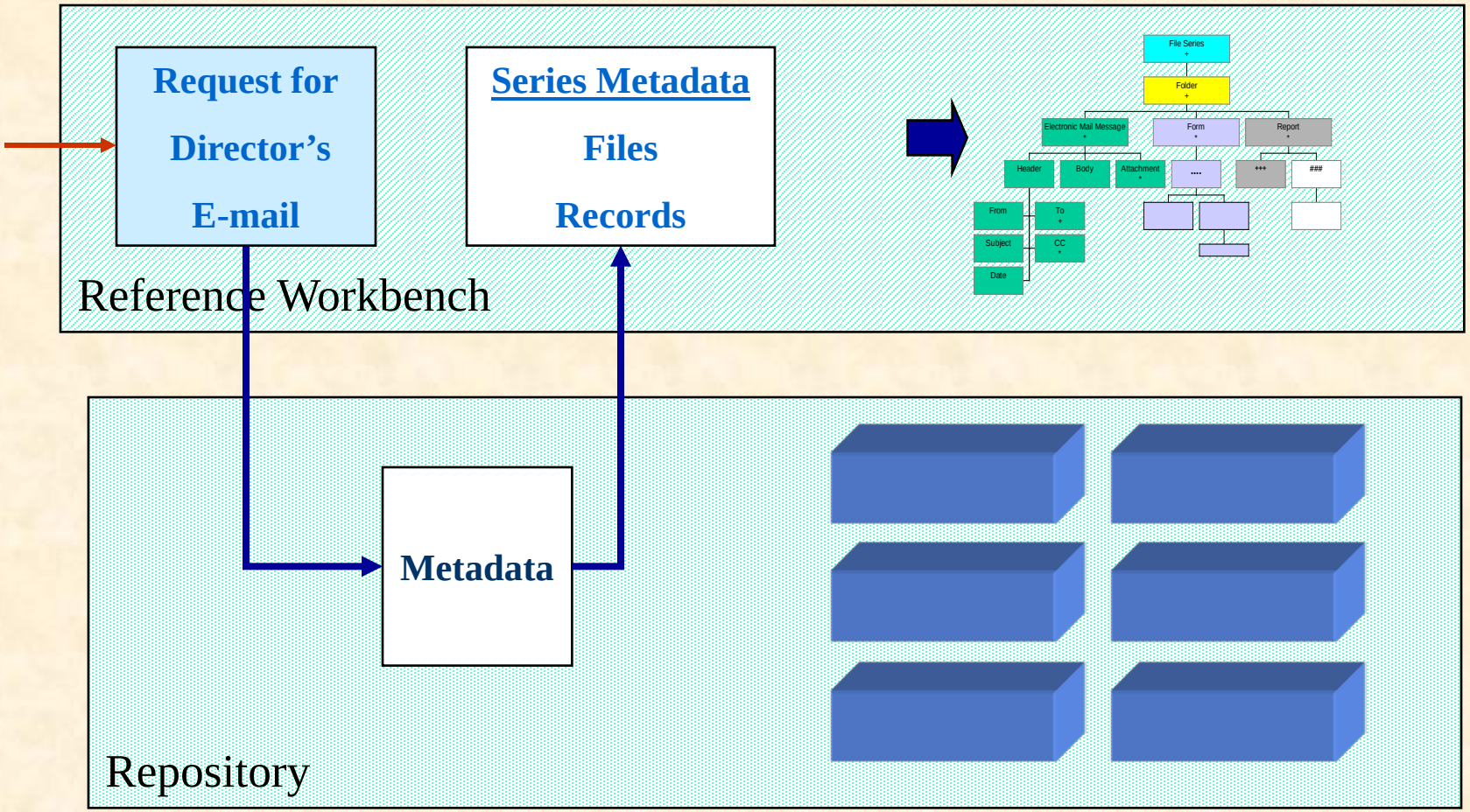
Risk Management: Multi-Valent Documents



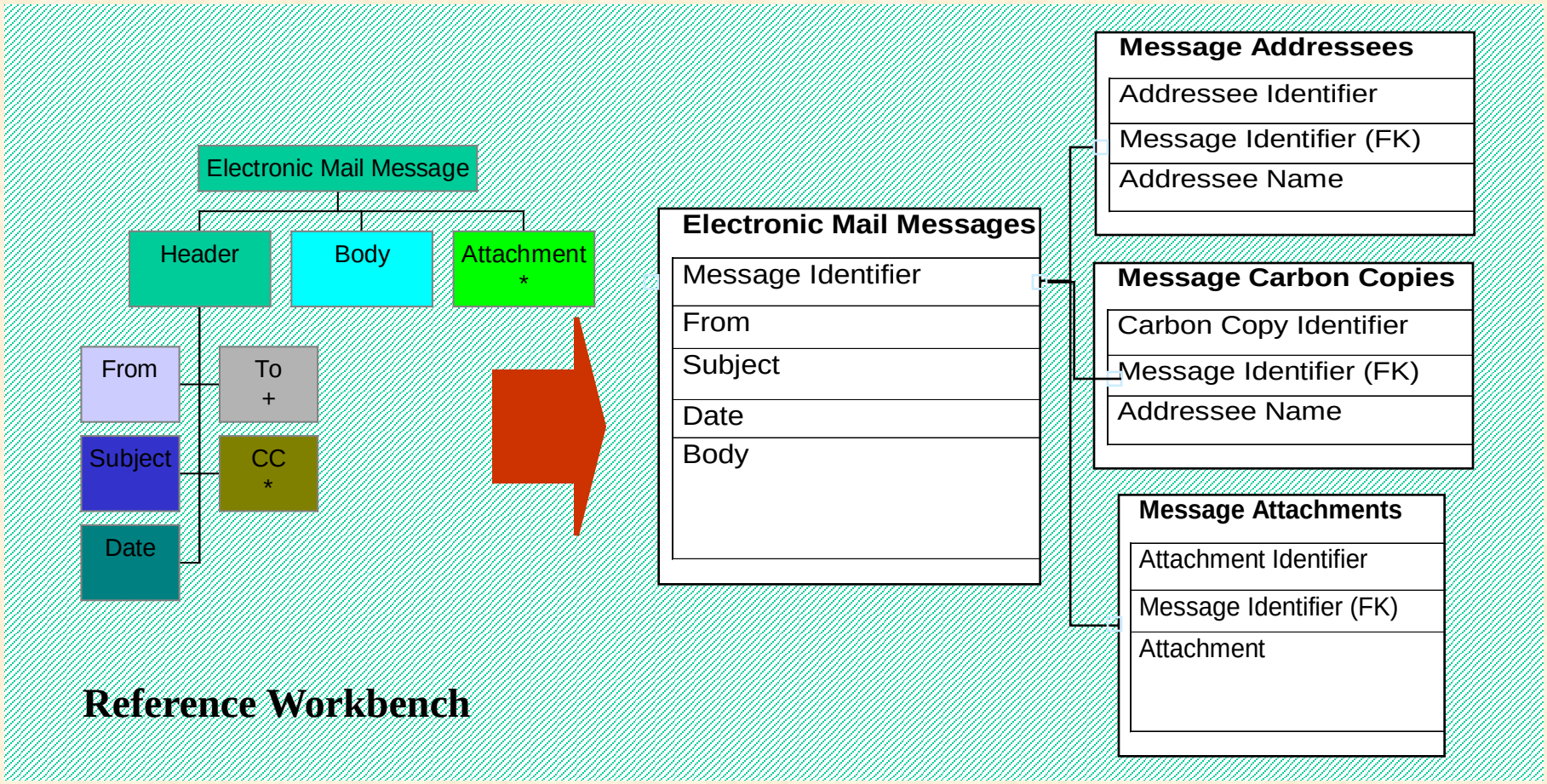
ERA: Reference Process



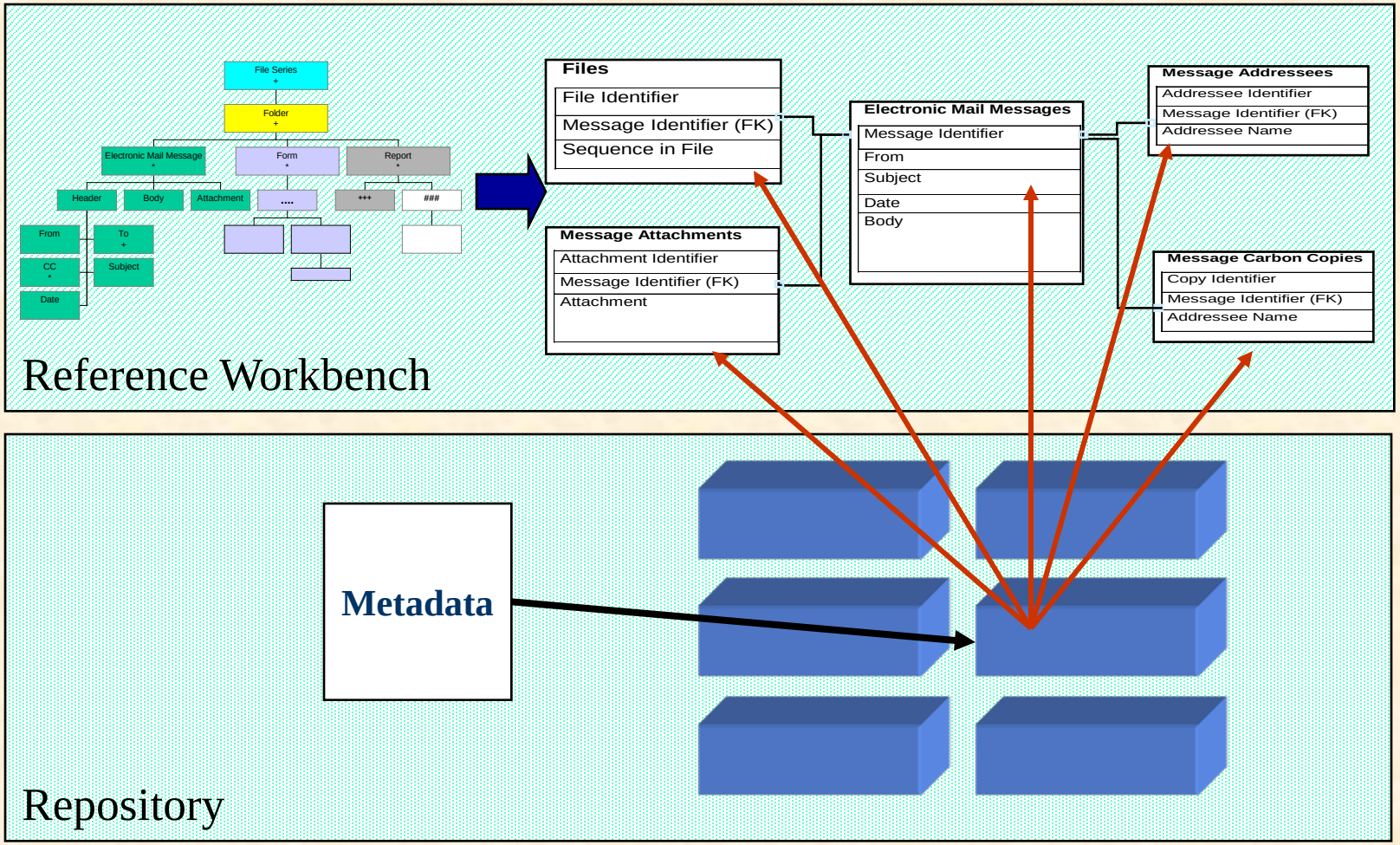
Process: **Check metadata for the series to identify relevant DTDs**



Translate E-mail DTD to Relational Database Structure



Process: **Retrieve the records and place in the target structure**



Persistent Object Preservation

- + Aims at independence of technological infrastructure
  - + Reduce threats to integrity and authenticity by minimizing changes over time.
- + Embeds changes in a comprehensive information management architecture designed for preservation
- + Inherently extensible
- + Facilitates use of future, advanced technologies, without requiring change in what is preserved.
- - Currently beyond state of the art of information technology.

Self-describing Objects for Records Management

- Facilitate management, exchange, and disposition of records
  - explicitly identify the content of records, files, series,...
  - express how content is organized
  - allow the content to be stored once and used in different documents
  - separate, but link, management of content and presentation
  - capture the relationships among documents and collections of documents
  - and support multiple views of a collection of documents
  - all in plain language

Thank you.

For more information:

www.nara.gov/era

Electronic Records Archives Program