A Blueprint for Representation Information in the OAIS Model

David Holdsworth

& Derek Sergeant

Leeds University, UK

`http://www.leeds.ac.uk/cedars/`

CEDARS Project

CEDARS Project

- Curl Examplars in Digital ARchiveS
  Collaborative project for libraries
  Funded by HEFCE/JISC

CEDARS Project

- Curl Examplars in Digital ARchiveS
  Collaborative project for libraries
  Funded by HEFCE/JISC

- *CAMiLEON*
  Collaborative project on emulation
  Funded by NSF/JISC
  Creative Archiving at Michigan and Leeds
  Emulating the Old on the New

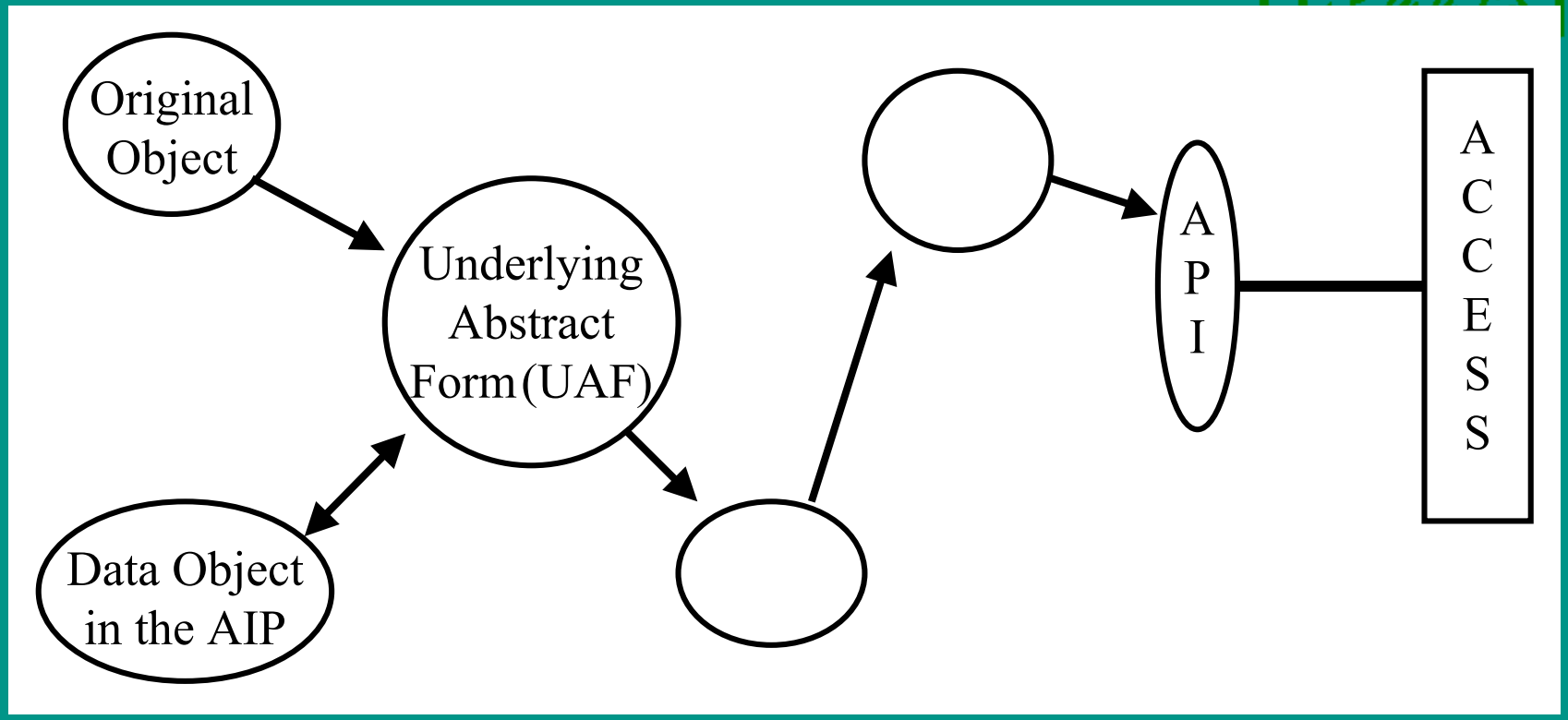
Current Status

- *Prototype Software to demonstrate Representation Nets*
- *AIP packaging (ASN.1)*
- *Meta-data design for PDI (XML - DTD)*

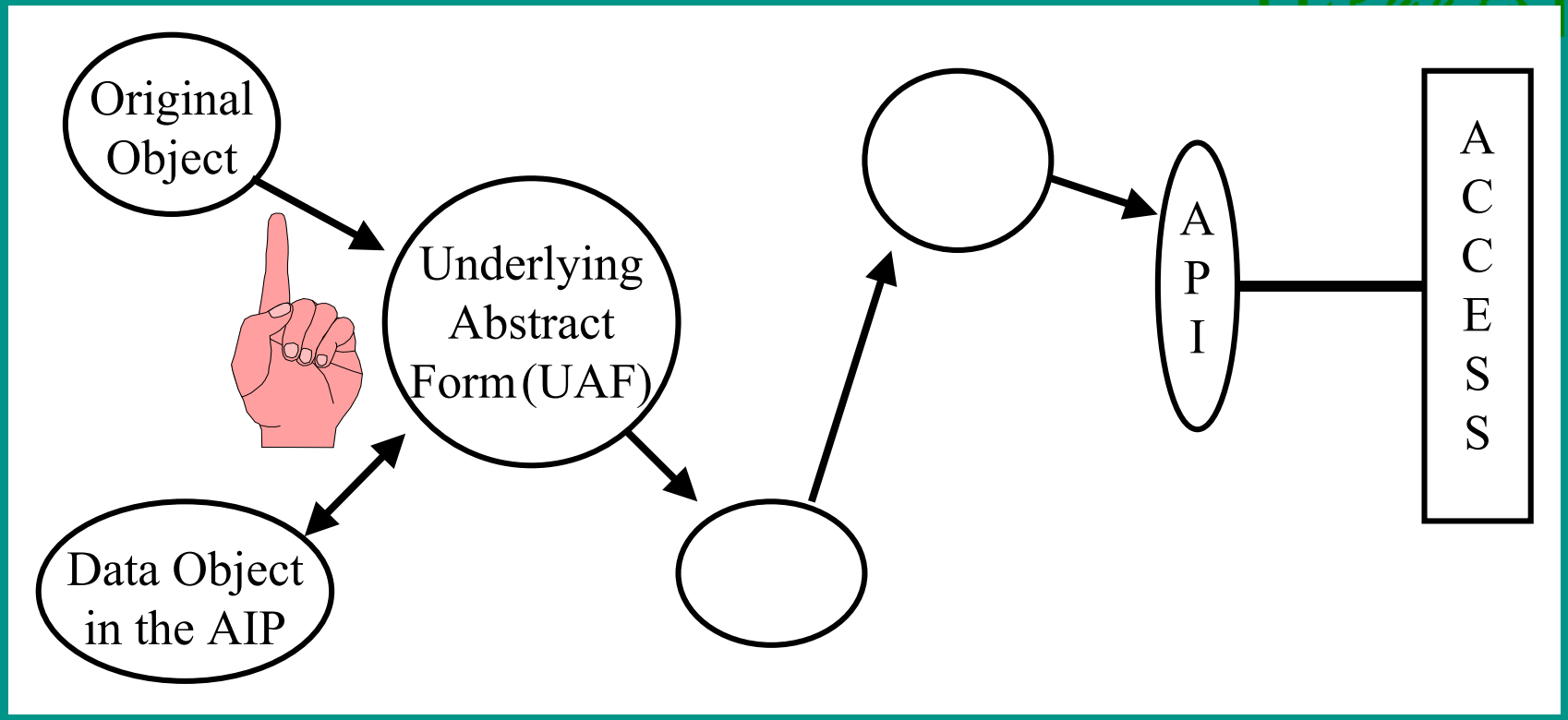
Why and Wherefore

- *Access to the* **intellectual content** *is the raison d'être of digital preservation*
- *A* **byte-stream** *can be stored for ever*

From Ingest to Access



Ingest 1

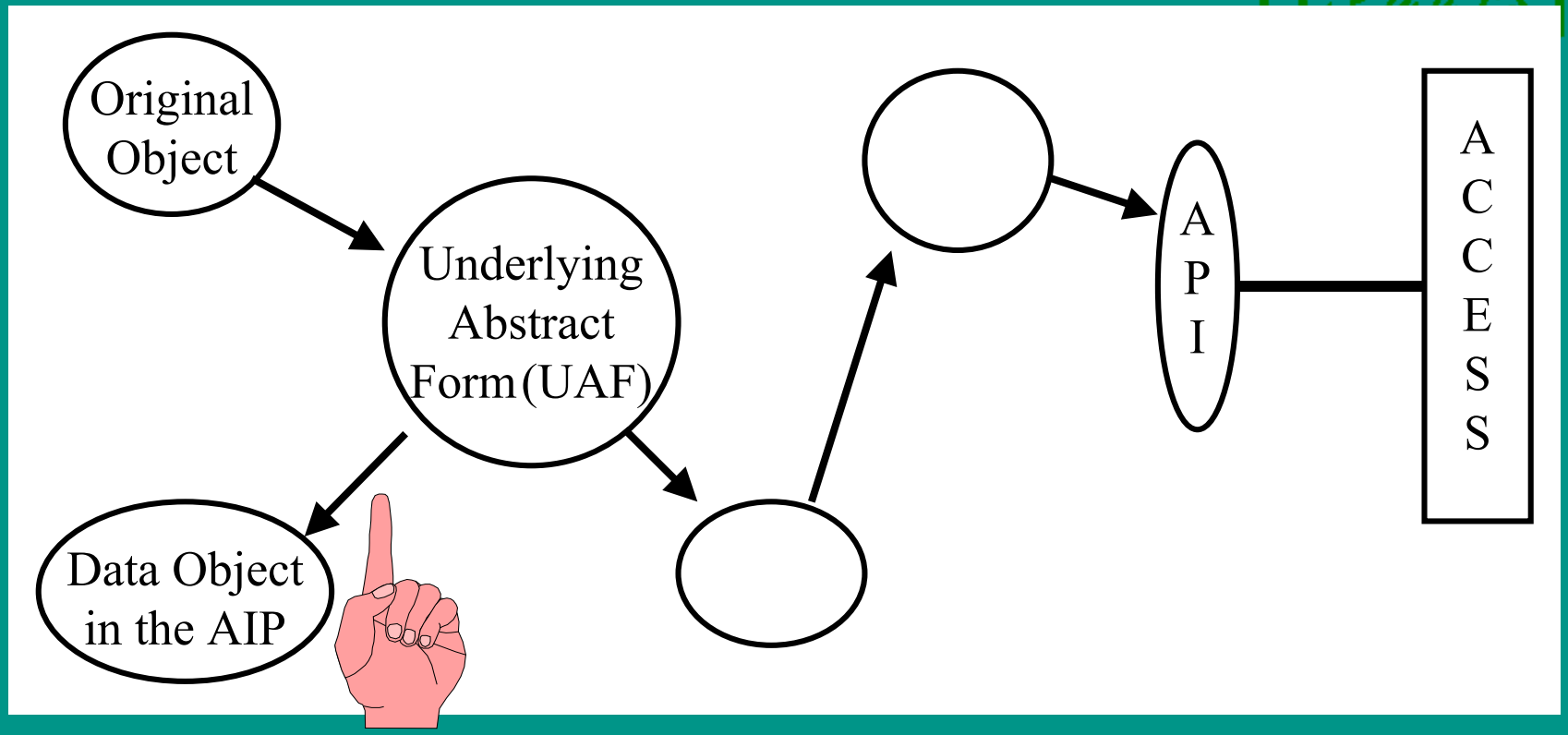


Detach the data from the medium

Underlying Abstract Form - UAF

- *The **UAF** is chosen to preserve the significant properties of the data set*
- *Identification of **significant properties** is vital*
- *At ingest the data set is mapped to a byte stream*

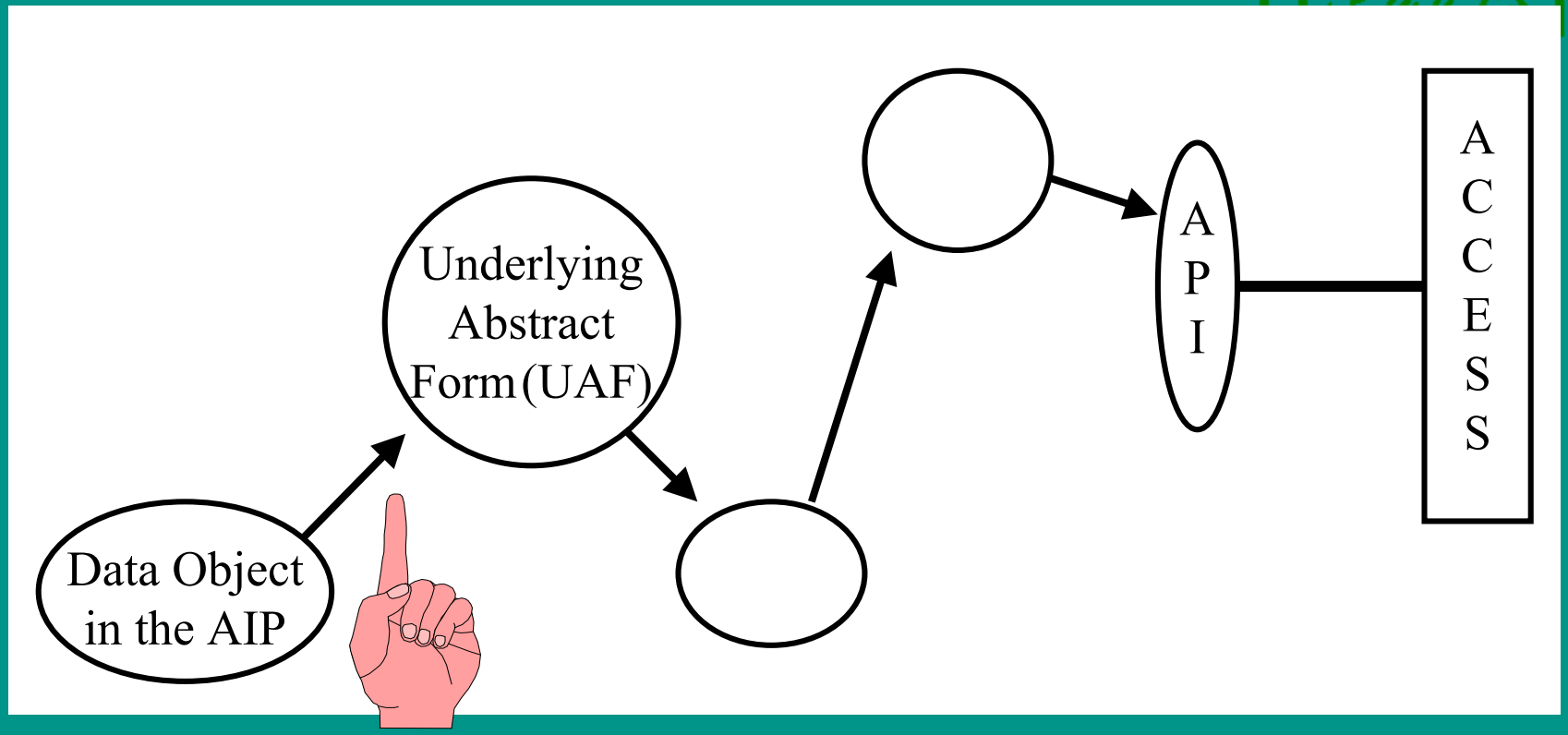
Ingest 2



Convert to a byte stream for long-term storage in an Archive Information Package (AIP)

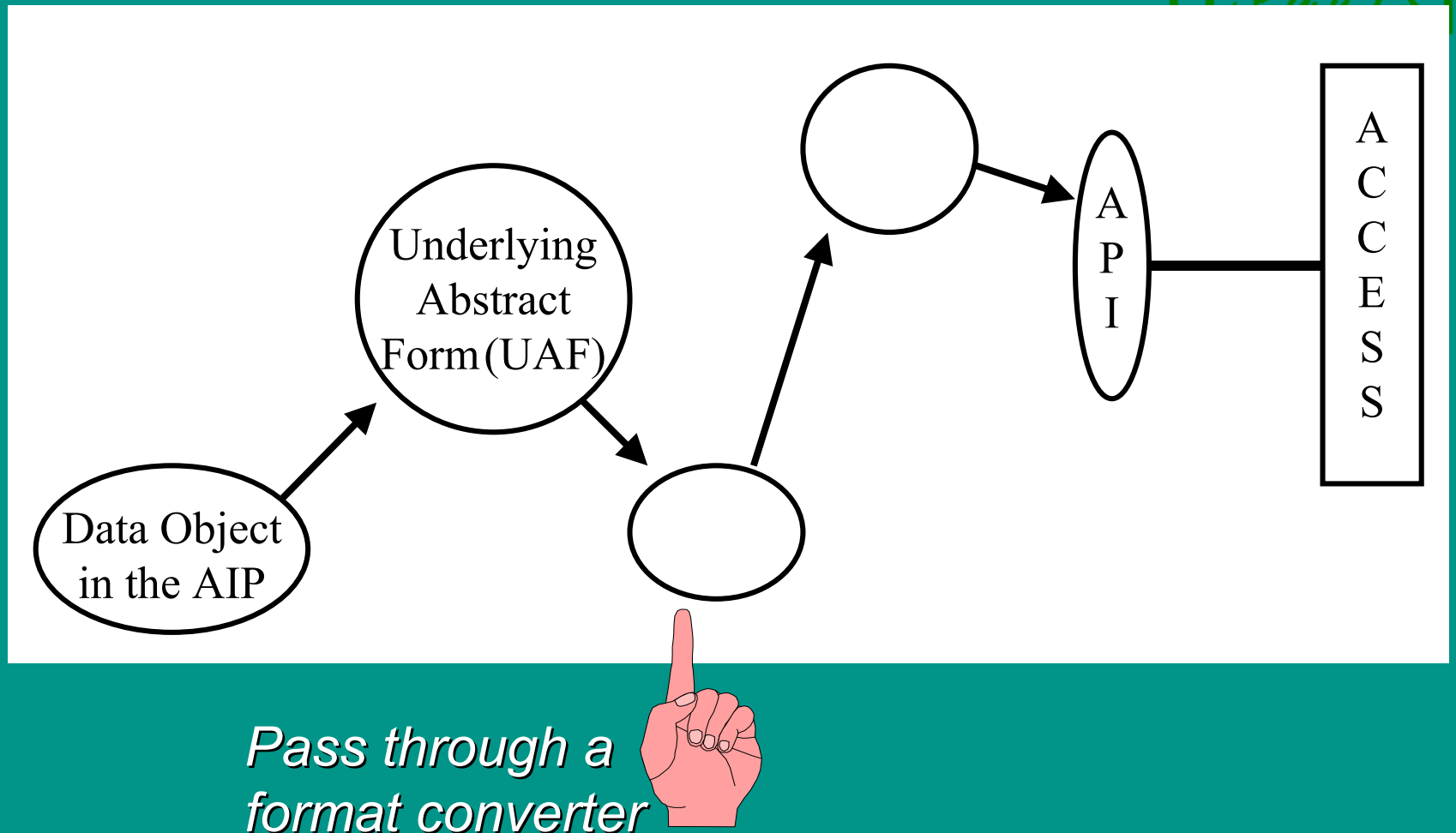
long-term storage

Access 1

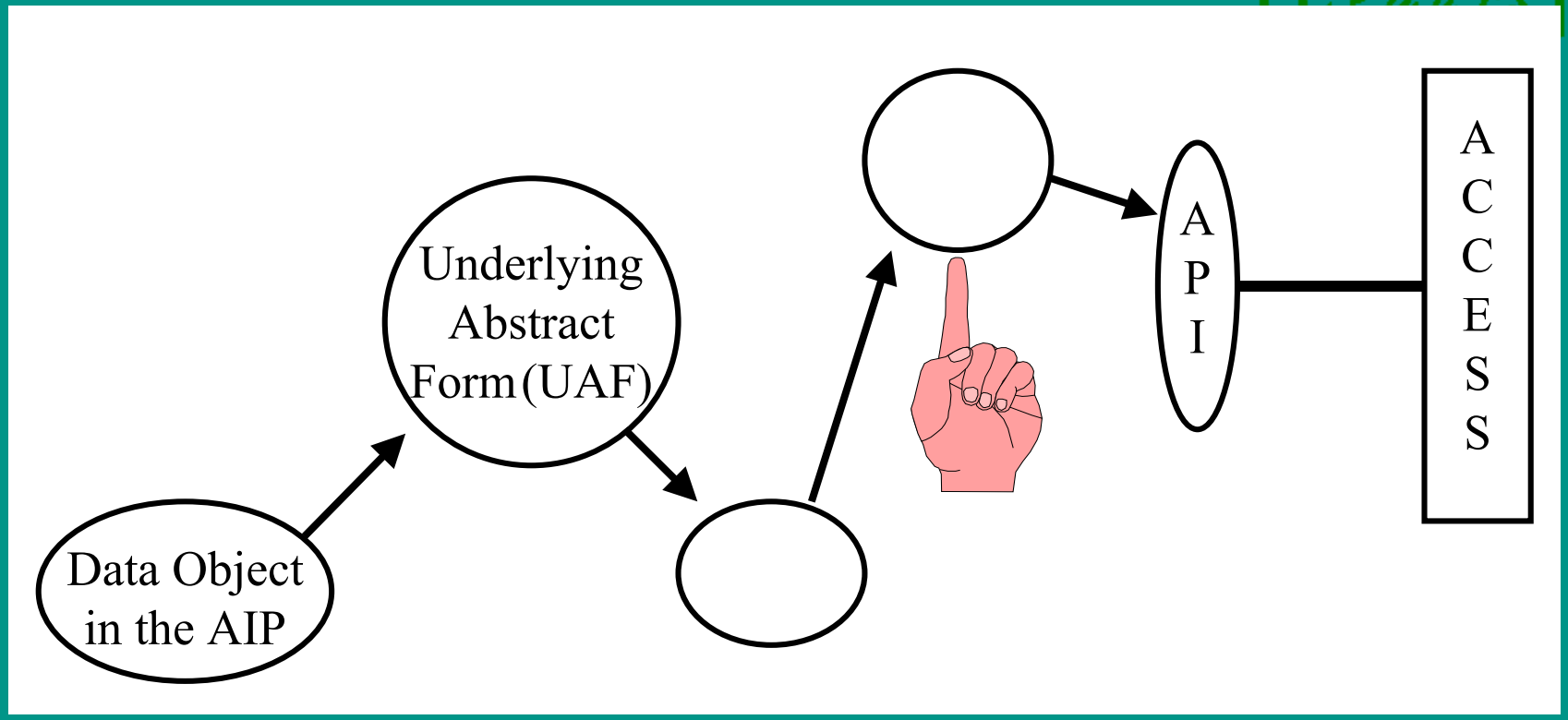


Rebuild the UAF

Access 2

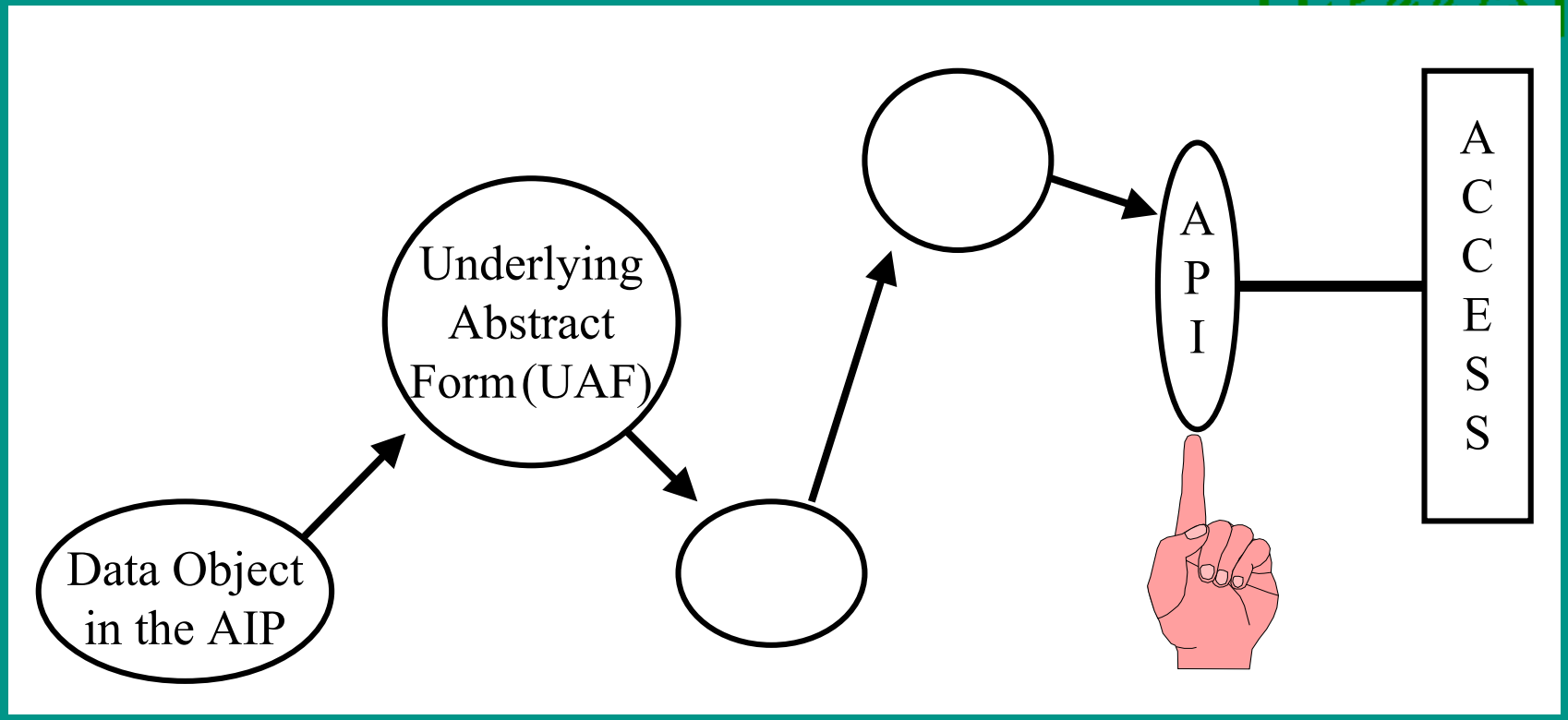

Underlying
Abstract
Form (UAF)
Data Object
in the AIP
API
ACCESS
Pass through a
format converter

Access 3



…… possibly many times ……..

Access 4



… to provide an API suitable for access to the intellectual content

Two Requirements

Two Requirements

- *ensure that you can find the preserved digital object*
  - *each preserved object must have a unique reference*

Two Requirements

- *ensure that you can find the preserved digital object*
  - *each preserved object must have a unique reference*
- *ensure that you can "understand" it when you've found it*
  - *understand is to be understood as a rather loose term for successful access to relevant aspects of the intellectual content.*

Cedars Reference ID - CRID

Cedars Reference ID - CRID

- *All references to the same object must be by quoting the same CRID*

Cedars Reference ID - CRID

- *All references to the same object must be by quoting the same CRID*
- *Resource discovery facilities can deliver CRIDs as results*

Cedars Reference ID - CRID

- *All references to the same object must be by quoting the same CRID*
- *Resource discovery facilities can deliver CRIDs as results*
- *Our Representation Net is held together by CRIDs*

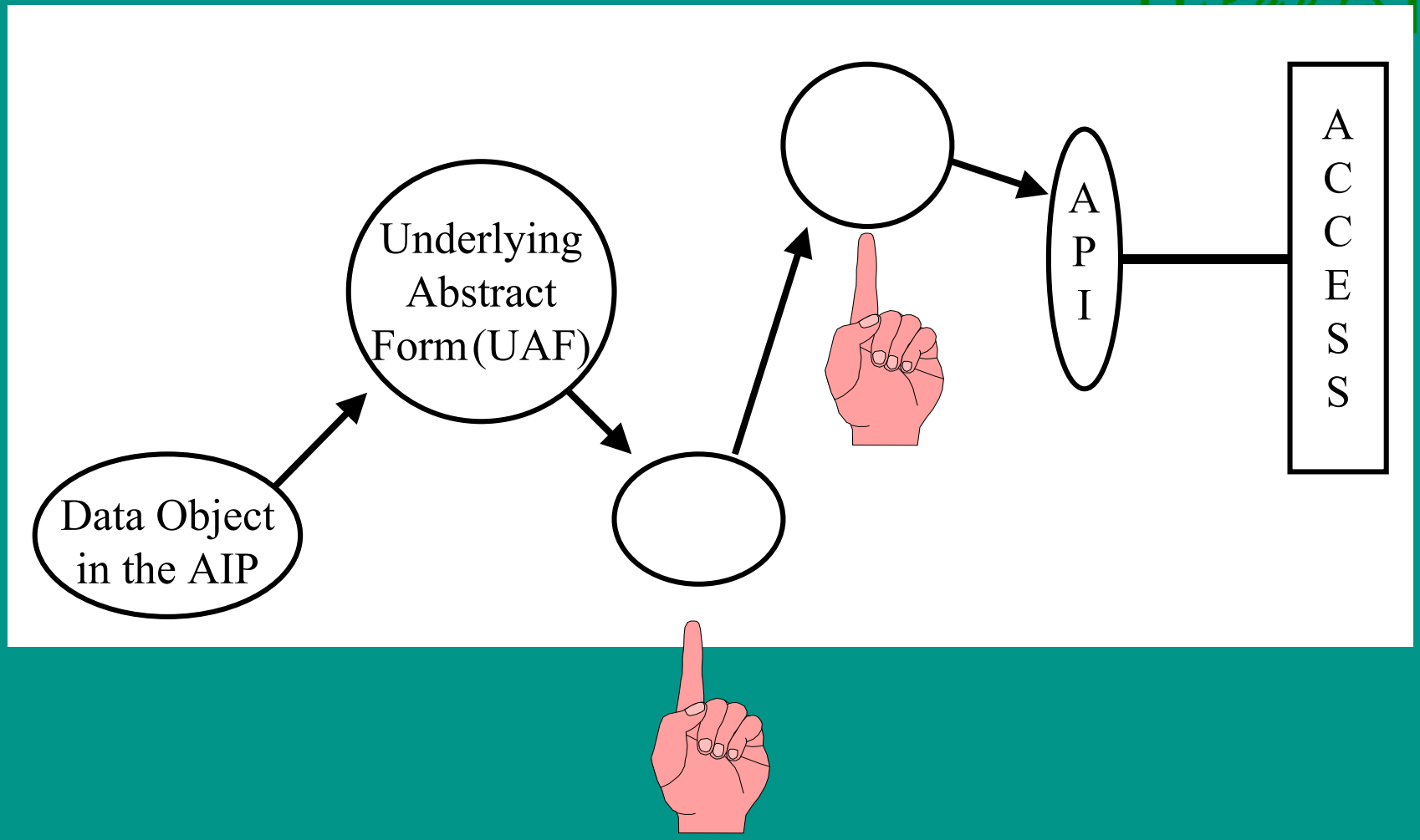
Name Allocation

- *multiple allocation authorities*
  - *c.f. Internet domains*
- *An allocation authority nominates a name server that knows the location of all its allocated CRIDs*
- *Each name server knows the identities of the allocation authorities' name servers*

Accessing Intellectual Content

- *Access to the intellectual content is the raison d'être of digital preservation*
- *A Render Analyse Engine (RAE) accesses data in one format and renders it in another*

Access via RAEs



Formats and RAEs

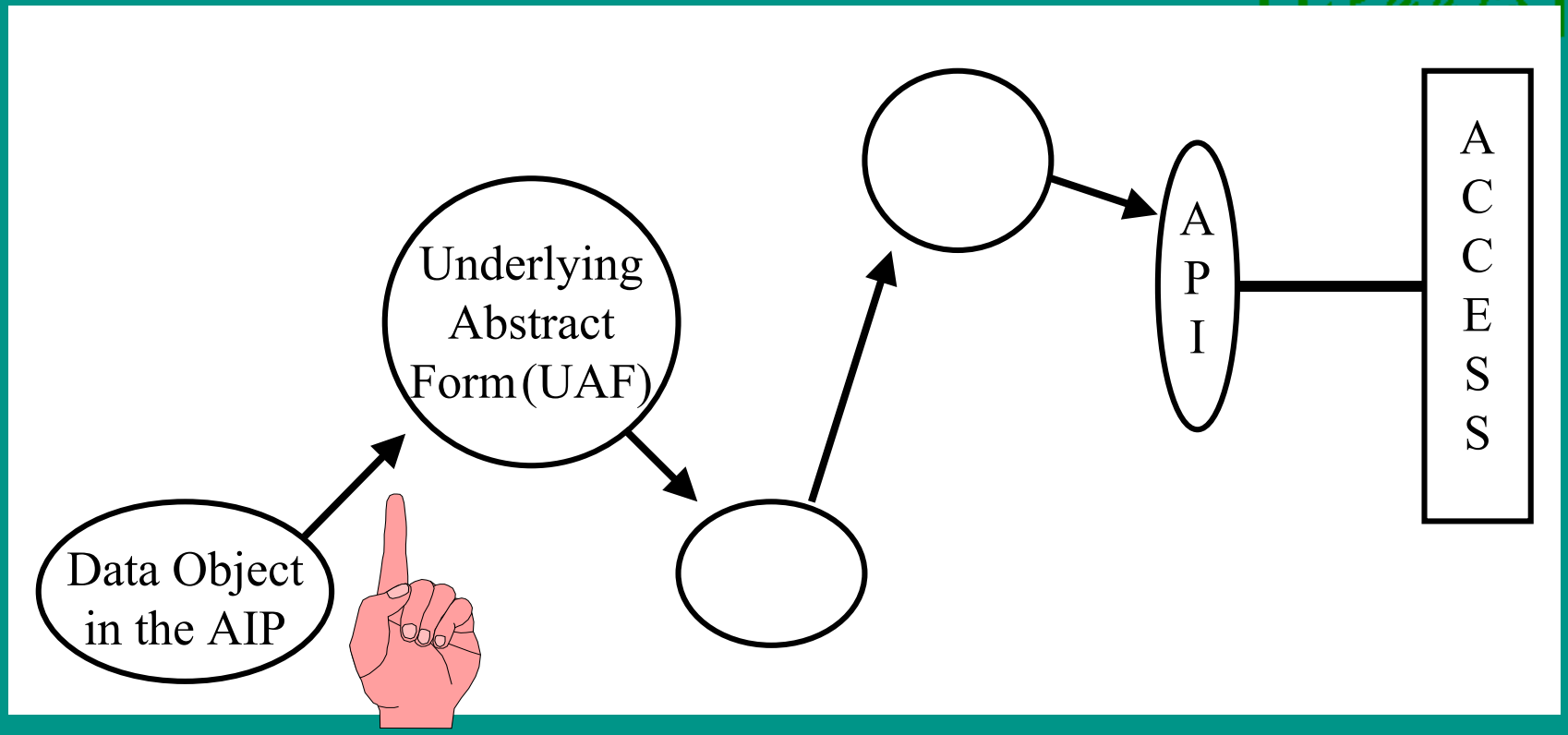
Formats and RAEs

- *An RAE converts from one format to another*

Formats and RAEs

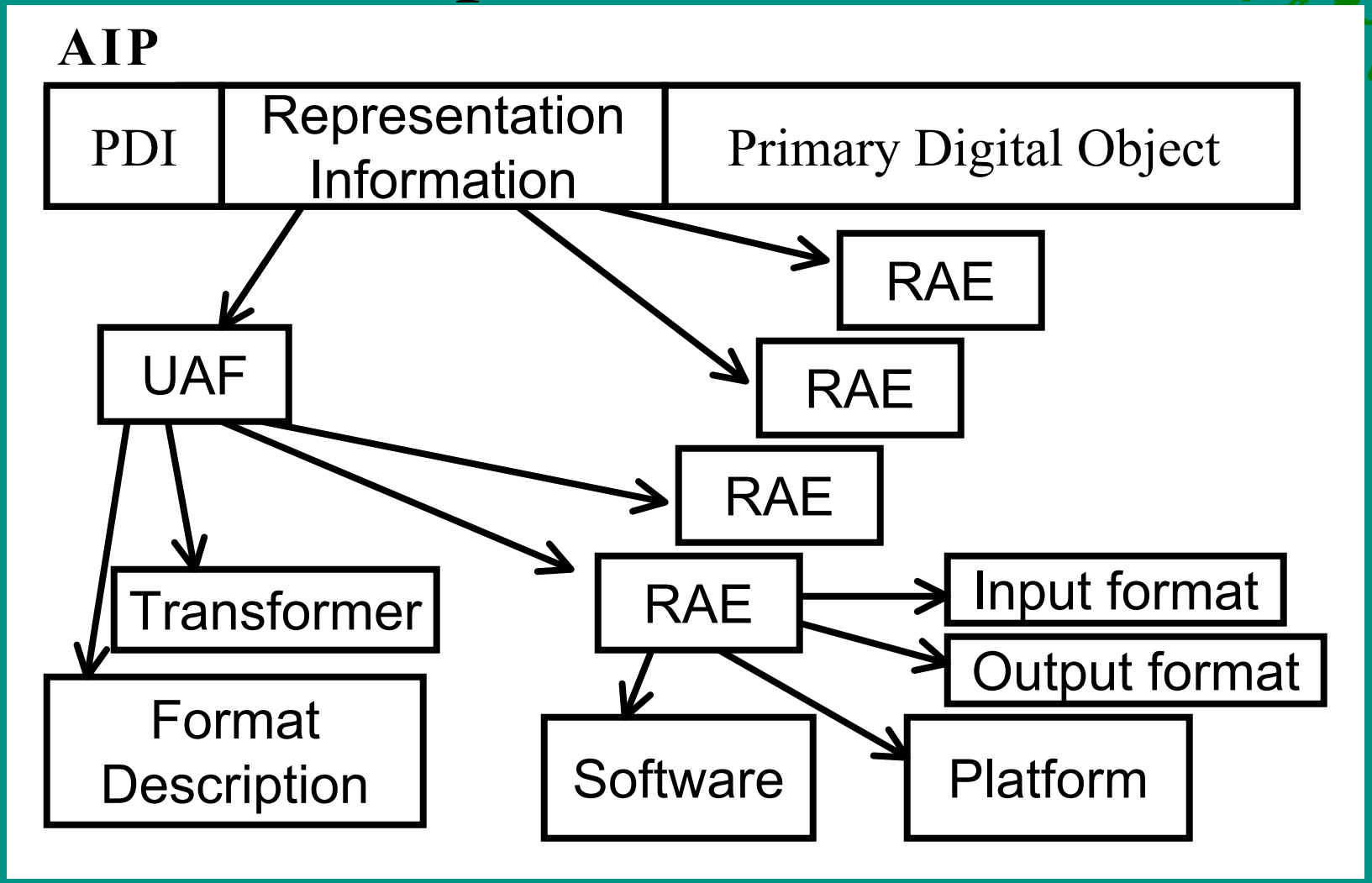
- *An RAE converts from one format to another*
- *A special RAE transforms the preserved byte stream into the UAF*

Access 1

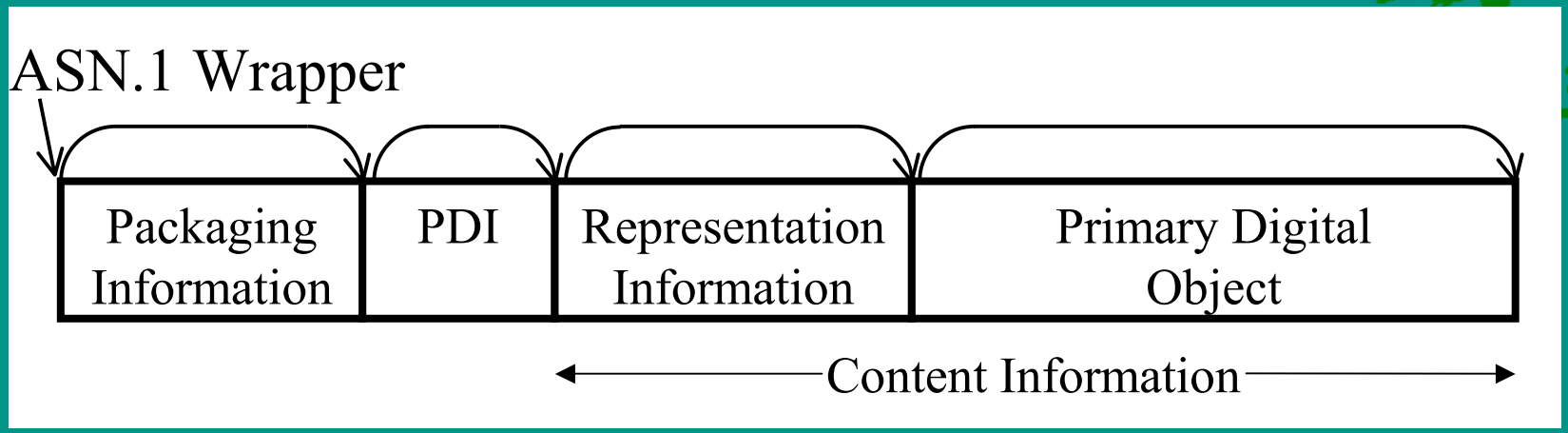


Rebuild the UAF from the byte stream

Cedars Representation Net



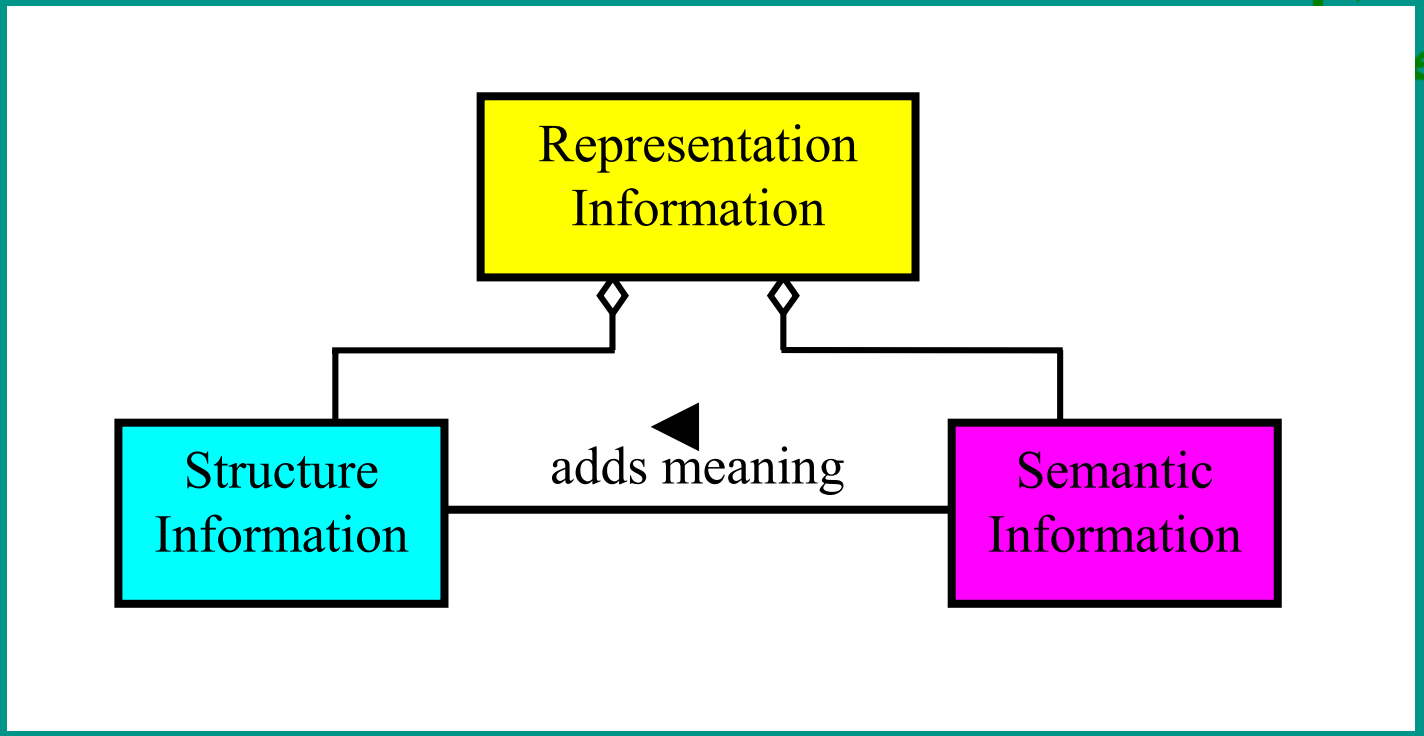
AIP - Implementation



There is an AIP for every object that has a CRID

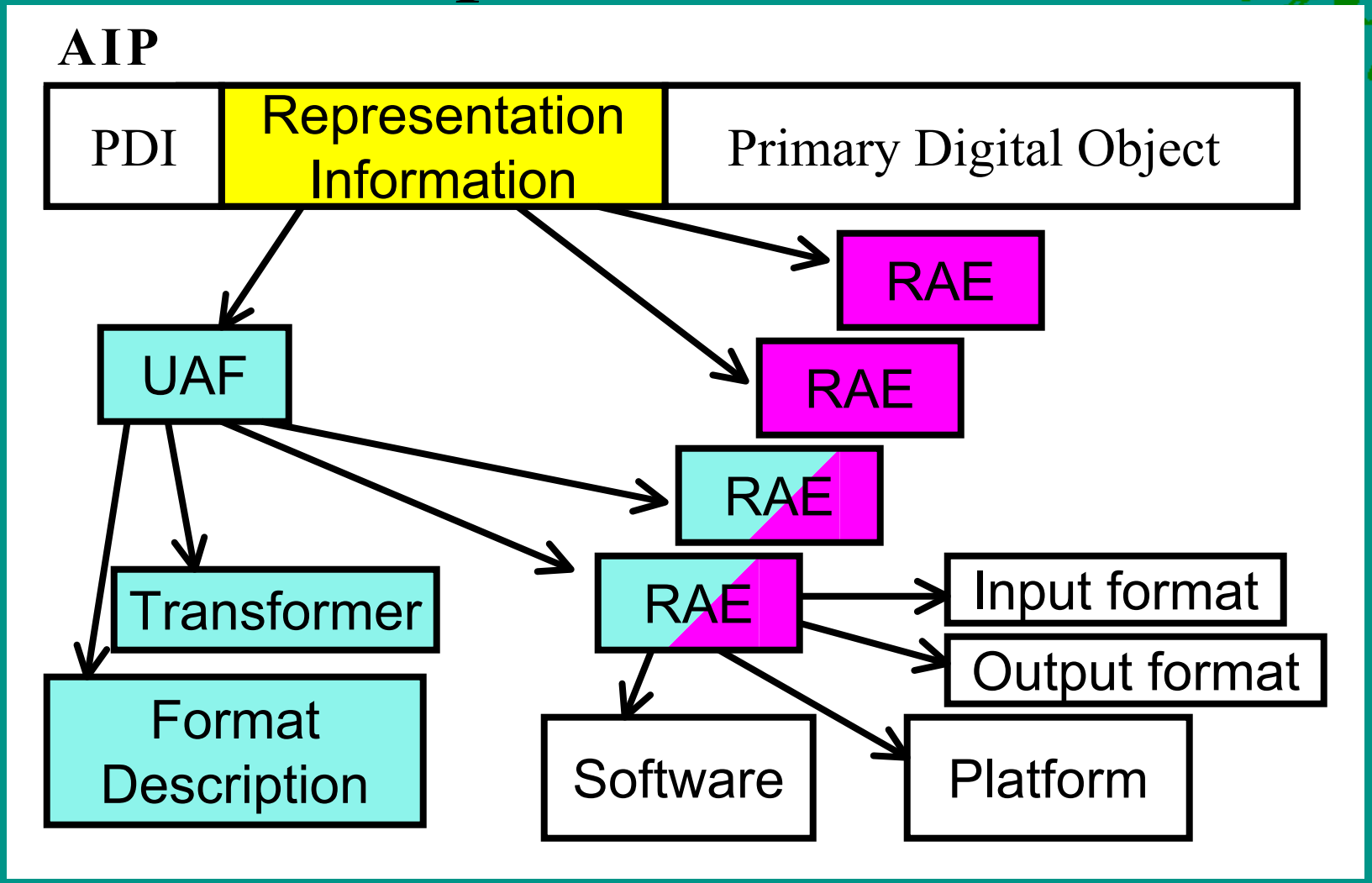
- **Primary digital object** = *the preserved object*
  - *the preserved byte stream*
- **RI** *needed to extract intellectual content*
- **PDI** = *Preservation Description Information*

OAIS Representation Information

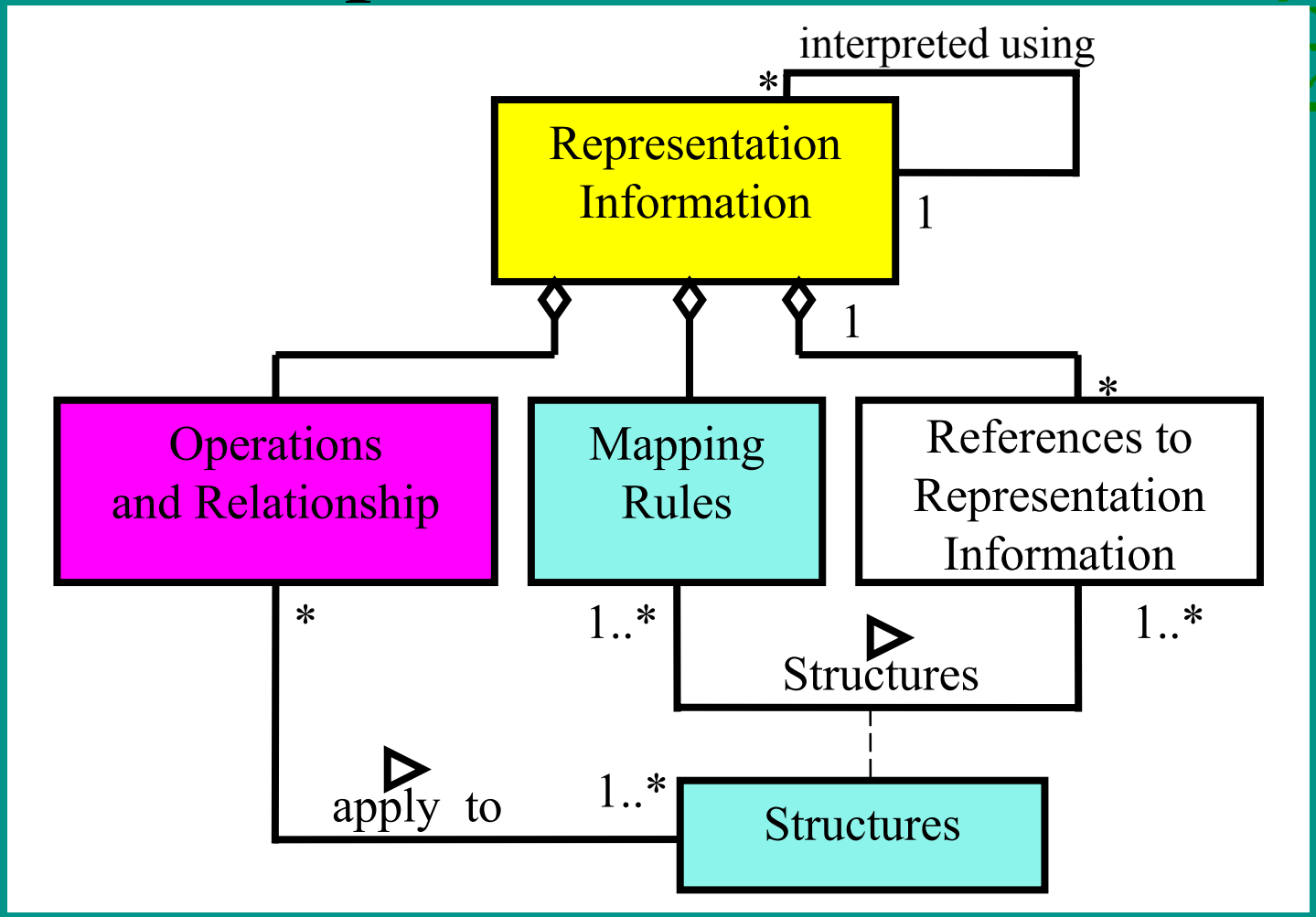


OAIS fig 4-10

Cedars Representation Net



OAIS Representation Nets



OAIS fig 4-11

Gödel's Theorem

Gödel's Theorem

- *Some representations (e.g. plain ASCII text, MS-WORD, HTML) are defined outside the system*

Gödel's Theorem

- *Some representations (e.g. plain ASCII text, MS-WORD, HTML) are defined outside the system*
- *All references to such a format are via the same CRID*

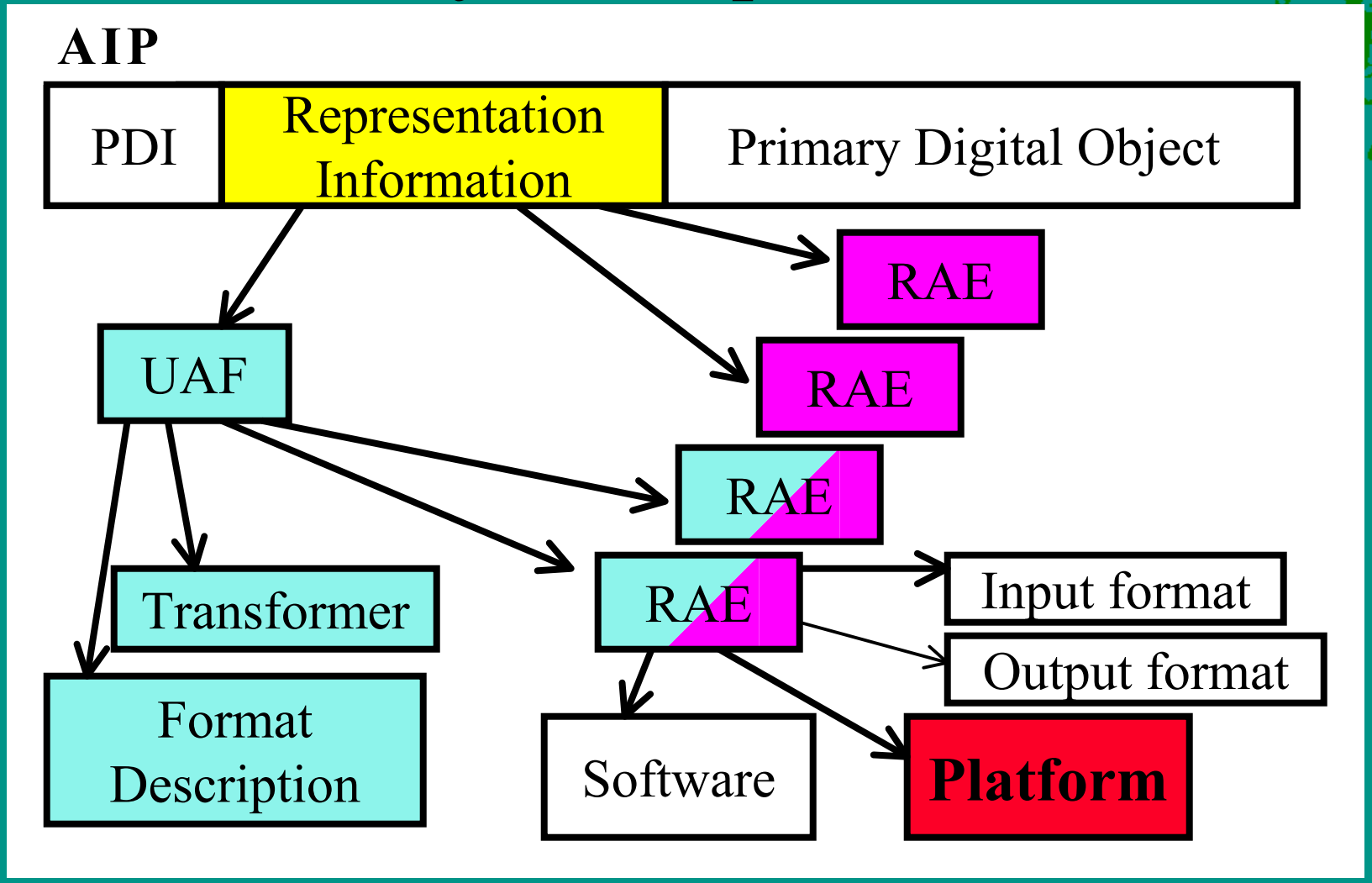
Gödel's Theorem

- *Some representations (e.g. plain ASCII text, MS-WORD, HTML) are defined outside the system*
- *All references to such a format are via the same CRID*
- *The ends of representation nets must be managed, to look out for obsolescence*

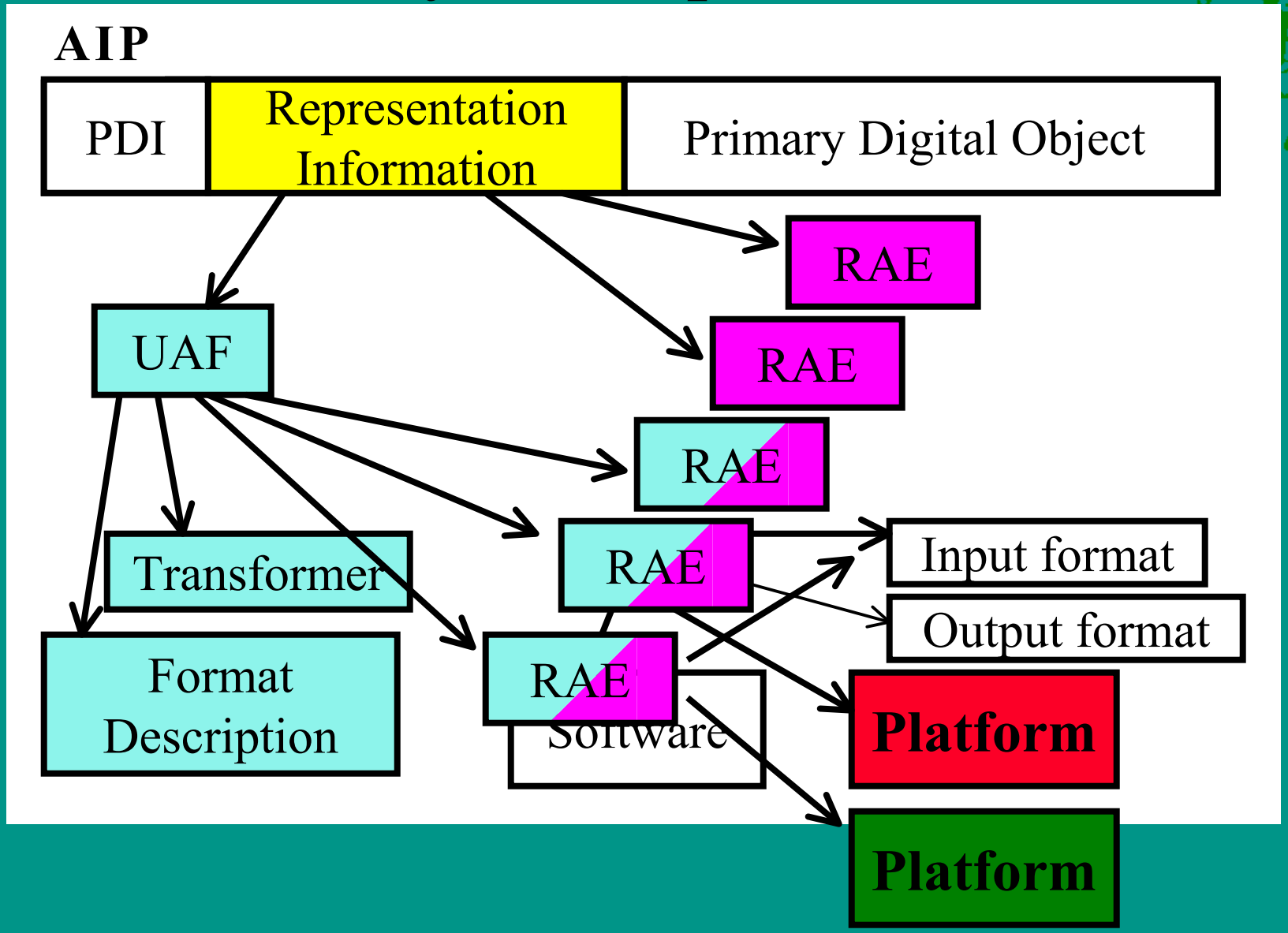
Gödel's Theorem

- *Some representations (e.g. plain ASCII text, MS-WORD, HTML) are defined outside the system*
- *All references to such a format are via the same CRID*
- *The ends of representation nets must be managed, to look out for obsolescence*
- *replace CRID destination with converter facility*

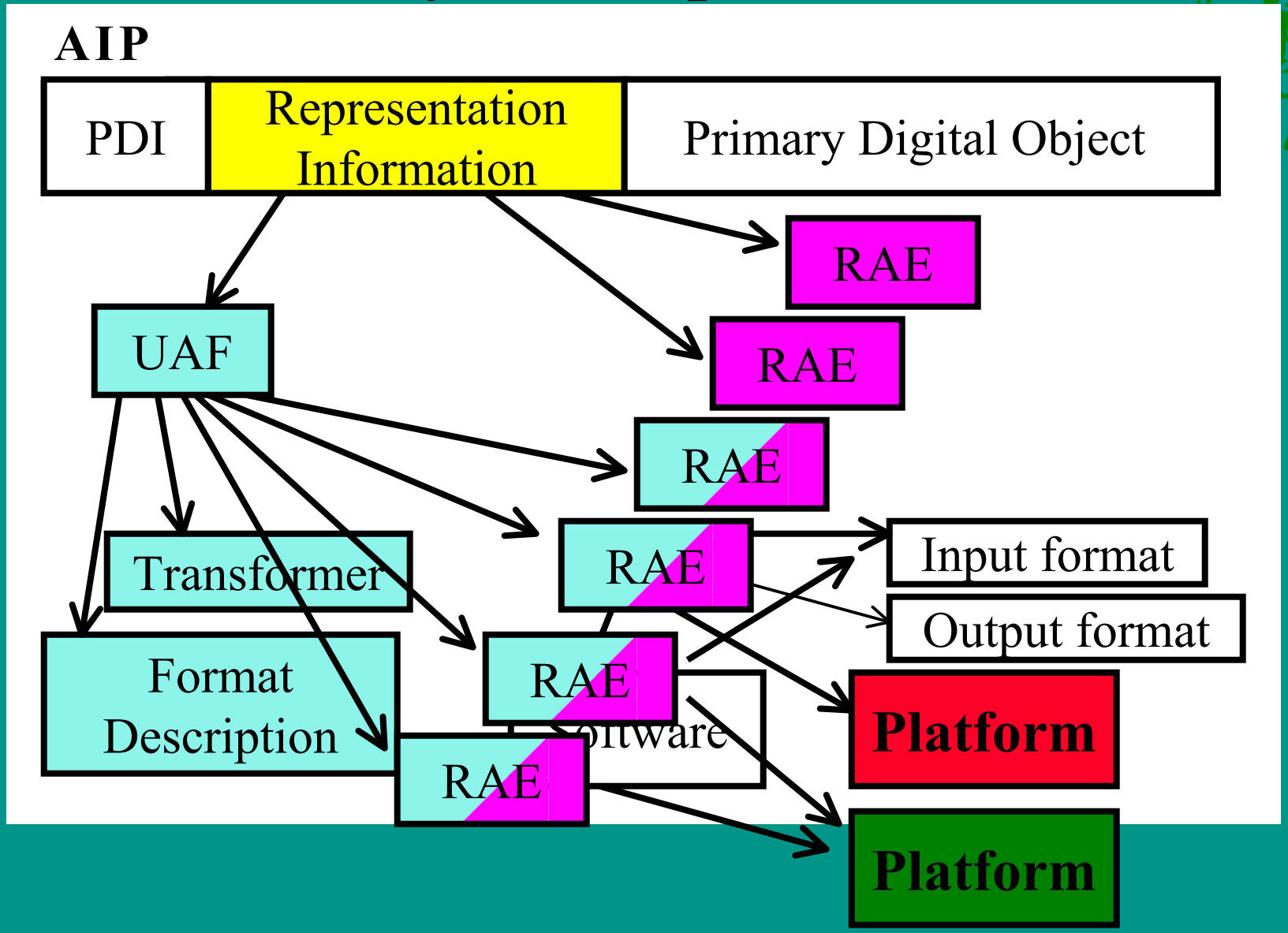
Evolution of the Representation Net



Evolution of the Representation Net



Evolution of the Representation Net



Obsolete data formats

Obsolete data formats

- *Keep the original byte-streams*

Obsolete data formats

- *Keep the original byte-streams*
- *Representation info leads to sofware capable of rendering the information*

Obsolete data formats

- *Keep the original byte-streams*
- *Representation info leads to sofware capable of rendering the information*
- *Archive management must lookout for dependence on rendering software that is about to become obsolete.*
  - *Can use software preservation techniques to preserve rendering sofware*

Emulation of Yesteryear

Emulation of Yesteryear

- *Today's desktop machine far exceeds the mainframe of the 1970s or even 80s*

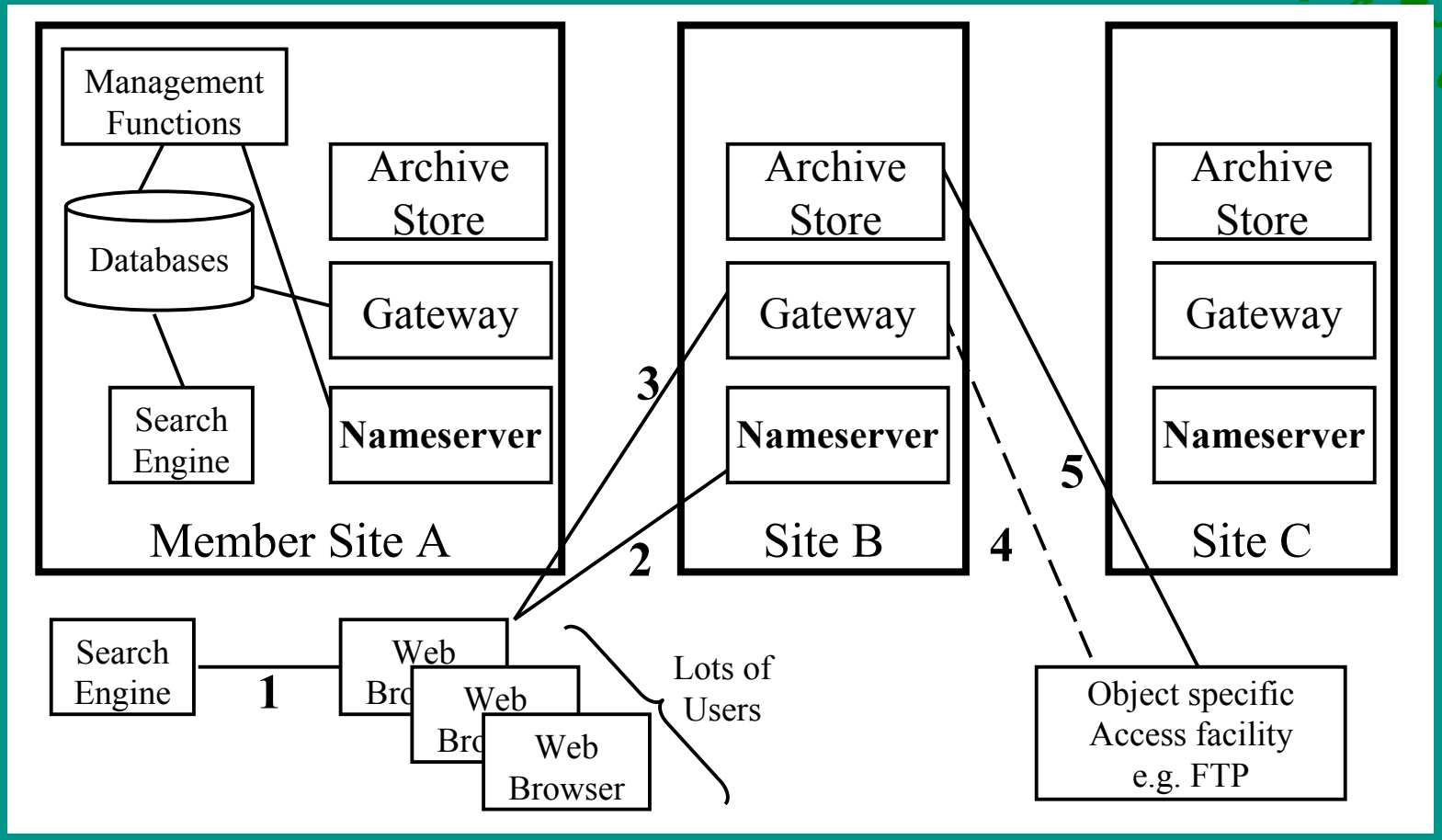
Emulation of Yesteryear

- *Today's desktop machine far exceeds the mainframe of the 1970s or even 80s*
- *George3 (1970s UK system)*
  - *Emulate the George3 executive*
    - *i.e. order code + system calls*

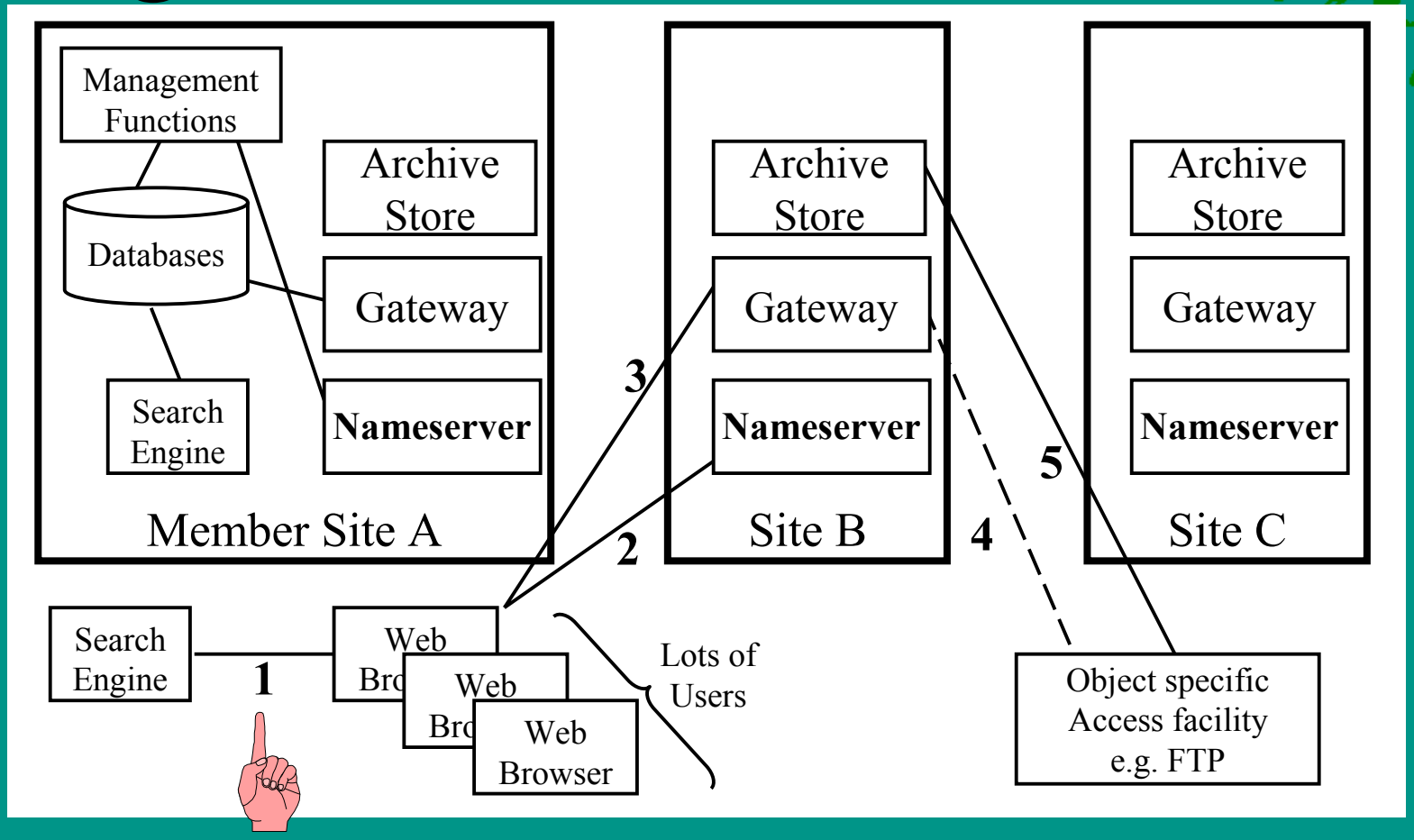
Emulation of Yesteryear

- *Today's desktop machine far exceeds the mainframe of the 1970s or even 80s*
- *George3 (1970s UK system)*
  - *Emulate the George3 executive*
    - *i.e. order code + system calls*
- *Constructing RI for obsolete materials proves a valuable test-bed for the model*

Distributed Architecture

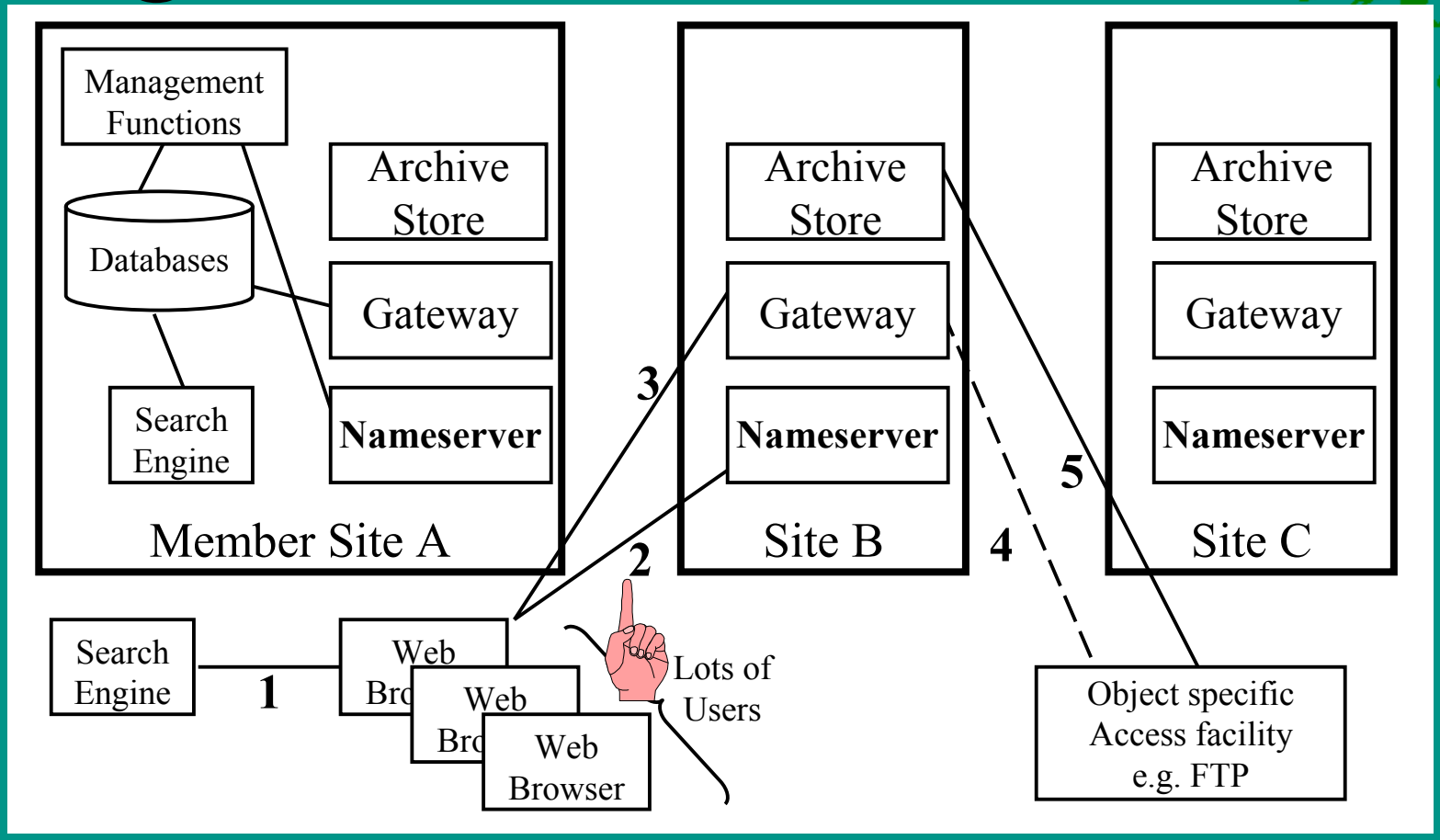


Stage 1



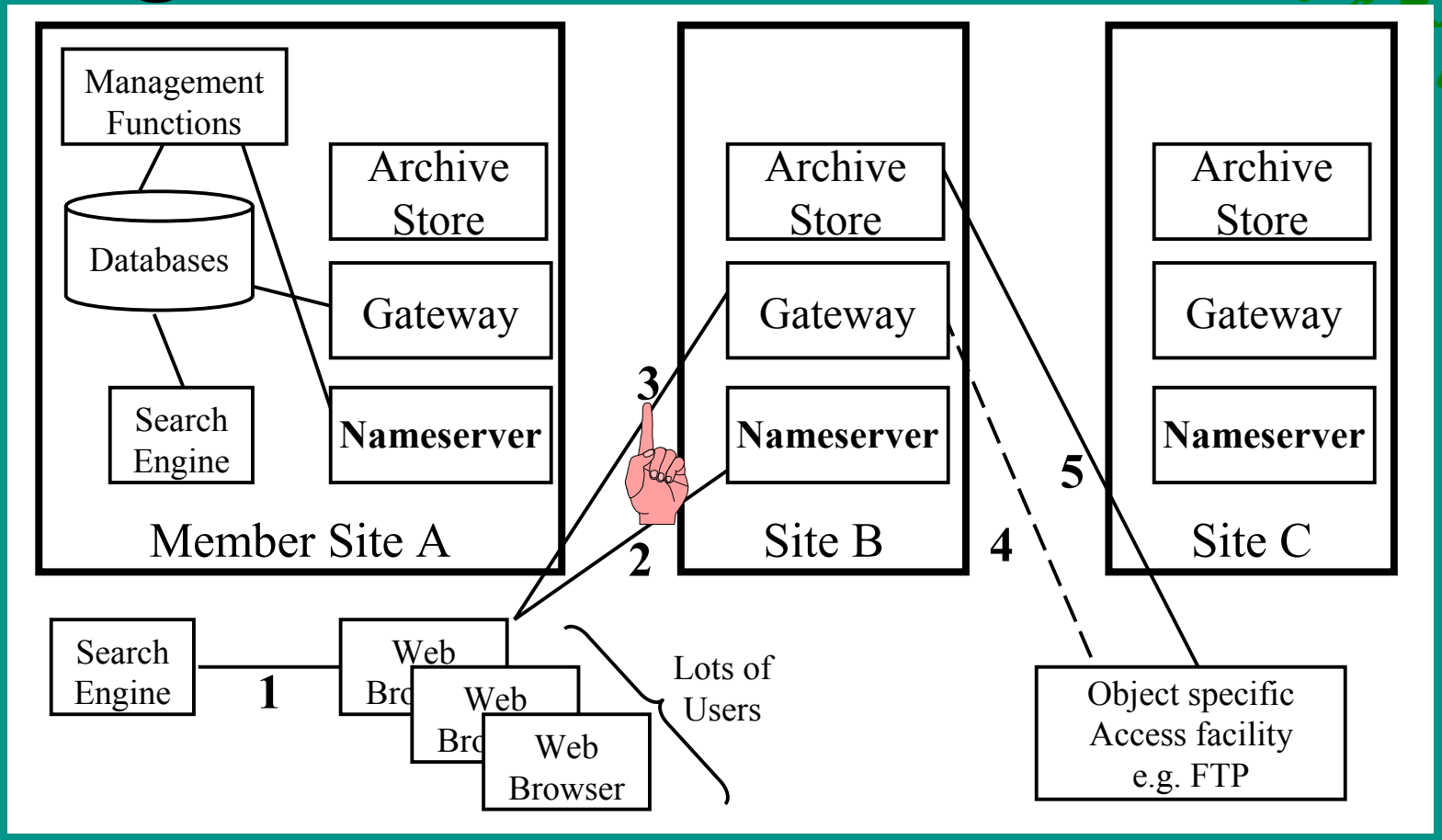
1 - Web interaction between search engine and end-user

Stage 2



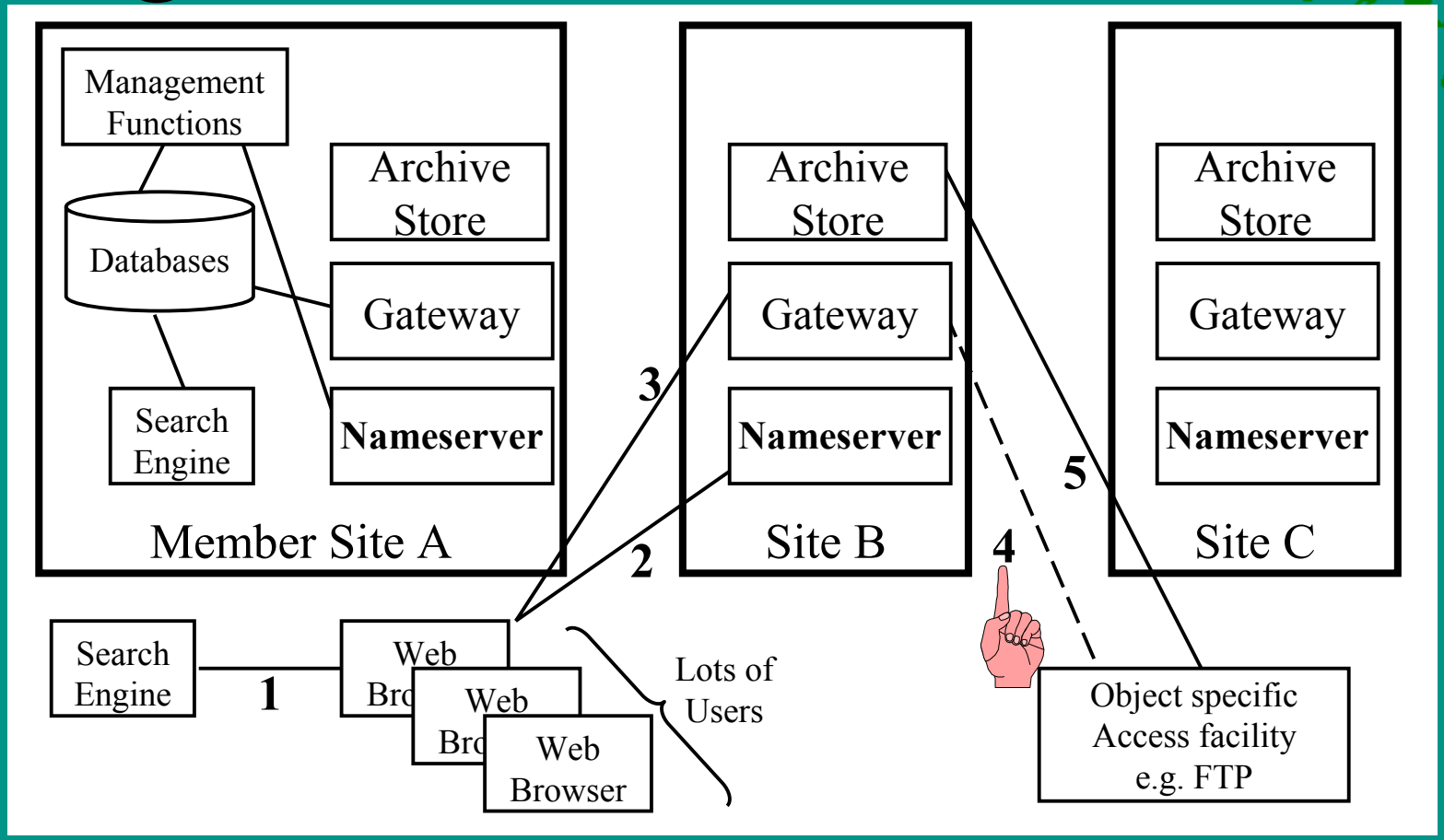
2 - HTTP call to the nameserver
redirects to the gateway to the desired object.

Stage 3



3 - Web interaction between gateway and end-user

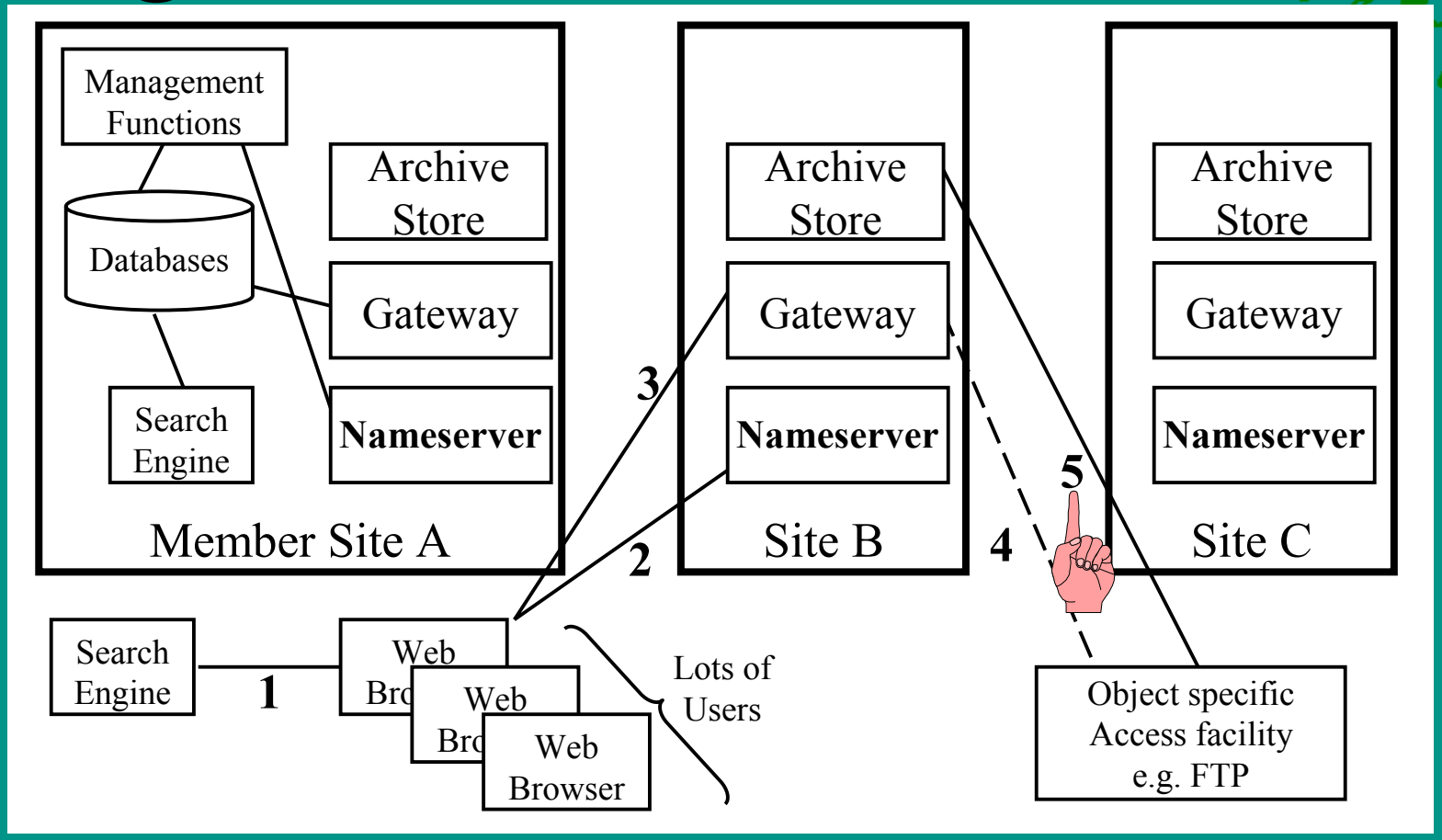
Stage 4



4 - response from gateway to end-user

may well be instructions to use FTP or to await a package delivered by mail (likely to include authentication)

Stage 5



5 - the digital object is delivered to the end-user

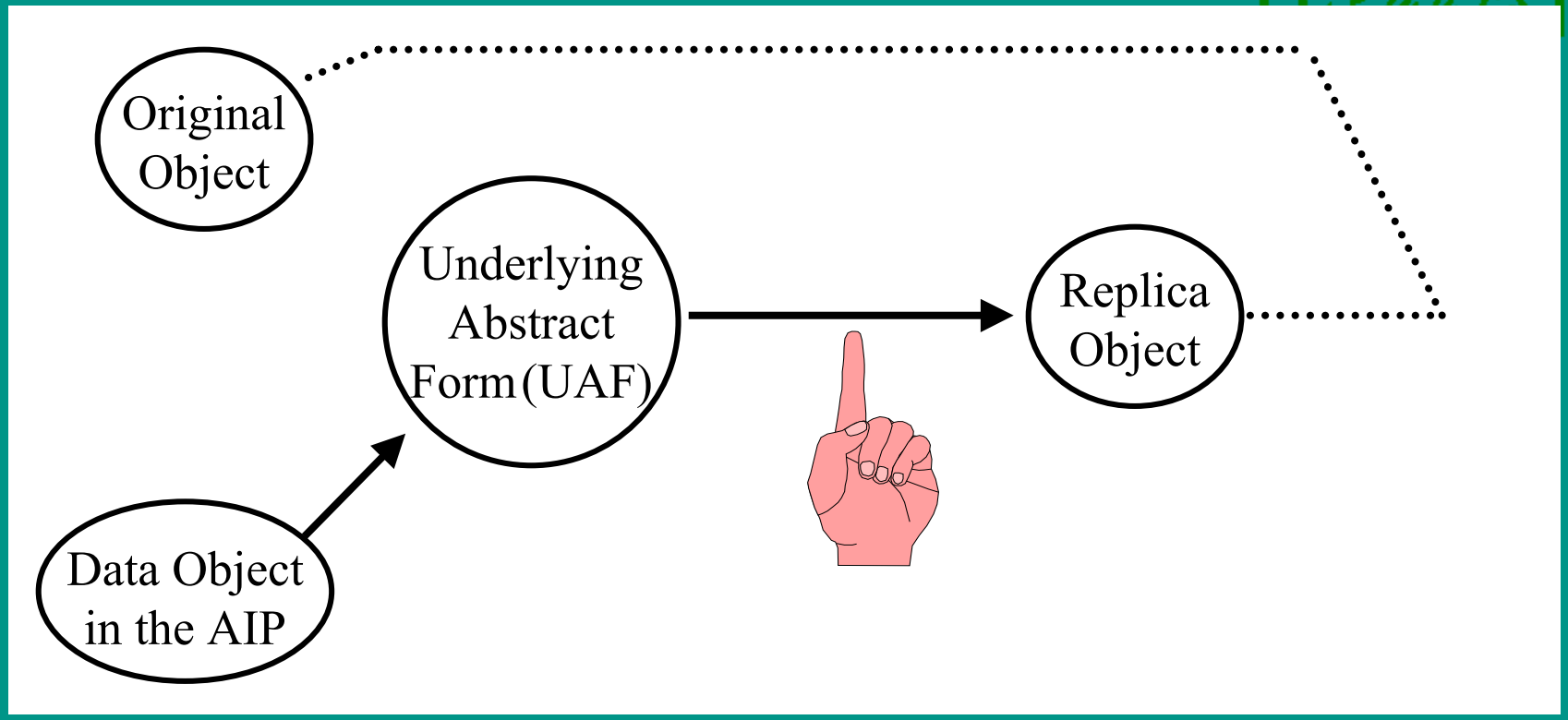
Representation Information enables interpretation

Choosing the UAF

- *A data set for preservation has an original existence in a storage system*
  - *e.g. file tree*
  - *PDF file*
- *The UAF is based on this original format*
- *Litmus test*

  Ability to recreate in principle

Litmus Test



Can recreate a "viable" replica from the UAF

Vital concepts

Vital concepts

- *A byte-stream can be stored for ever*
  - *Complex data streams must be mapped into byte-streams, and mapped back again for use.*

Vital concepts

- *A byte-stream can be stored for ever*
  - *Complex data streams must be mapped into byte-streams, and mapped back again for use.*
- *Representation Information preserves access to intellectual content*
  - *makes emulation possible*

Vital concepts

- *A byte-stream can be stored for ever*
  - *Complex data streams must be mapped into byte-streams, and mapped back again for use.*
- *Representation Information preserves access to intellectual content*
  - *makes emulation possible*
- *Gödel Ends are monitored for obsolescence*

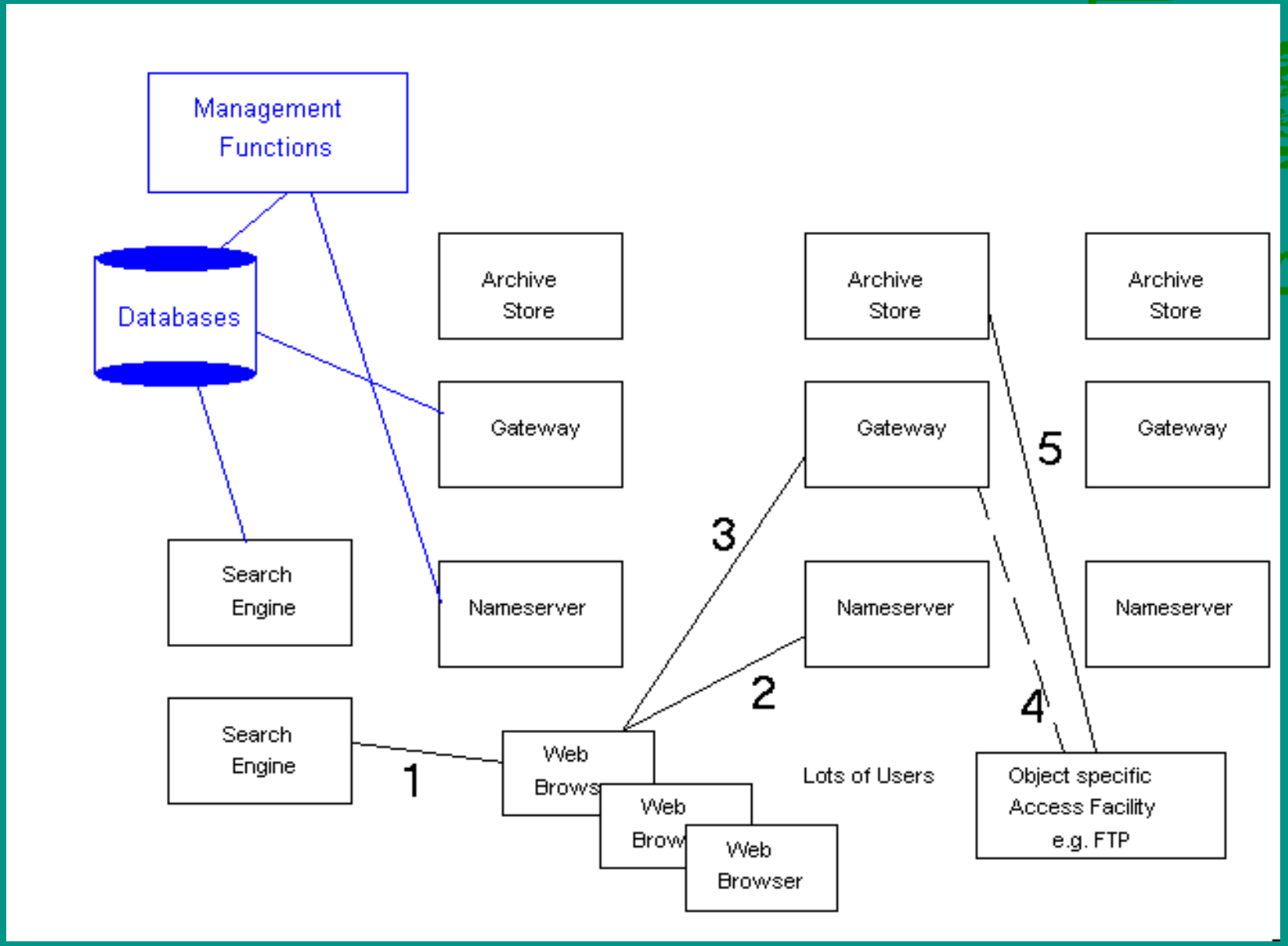
A Blueprint for Representation Information in the OAIS Model

David Holdsworth

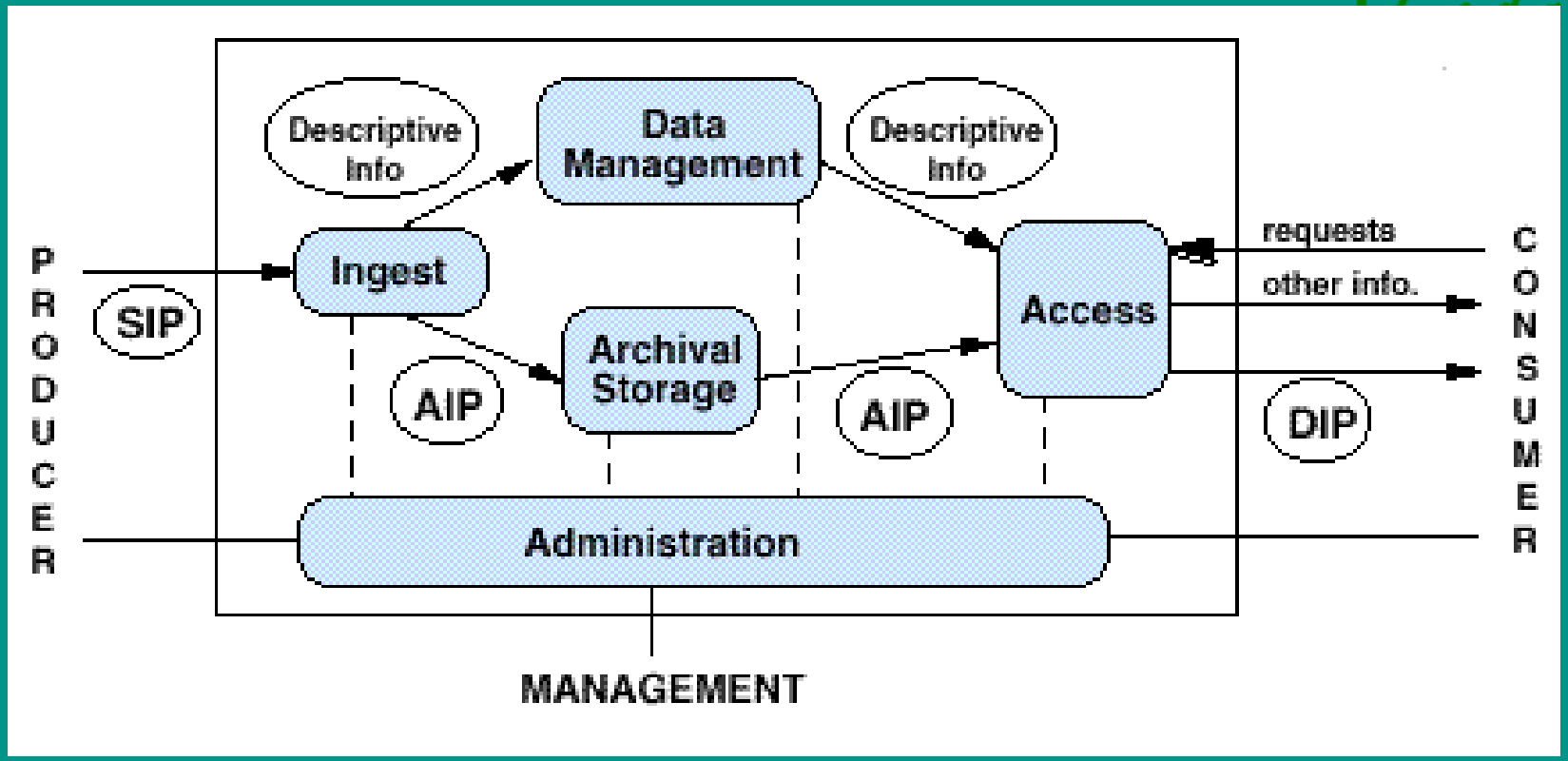
& Derek Sergeant

Leeds University, UK

`http://www.leeds.ac.uk/cedars/`


Management Functions
Databases
Archive Store
Gateway
Nameserver
Search Engine
Search Engine
Web Brows
Web Brow
Web Browser
1
2
3
4
5
Archive Store
Gateway
Nameserver
Lots of Users
Archive Store
Gateway
Nameserver
Object specific Access Facility e.g. FTP

OAIS Model


Descriptive Info
Data Management
Descriptive Info
PRODUCER
SIP
Ingest
AIP
Archival Storage
AIP
Access
requests
other info.
DIP
CONSUMER
Administration
MANAGEMENT