# Preservation of Real-World Data: The Case for Preservation DataStores

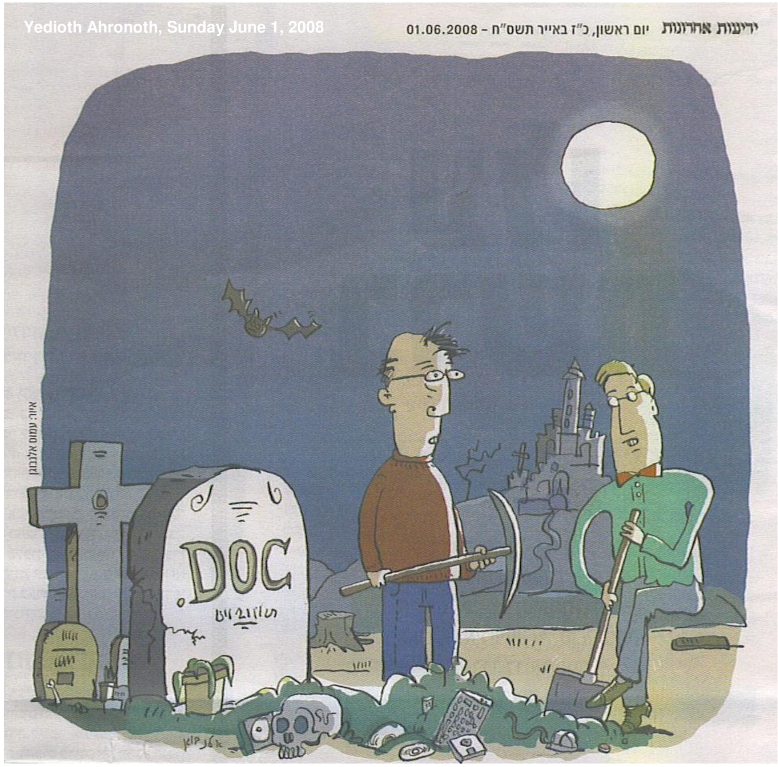
Simona Cohen, **Michael Factor**, Dalit Naor,
Leeat Ramati, Petra Reshef, Shahar Ronen

IBM Haifa Research Labs

http://www.haifa.il.ibm.com/projects/storage/ltdp/index.shtml

# What is Long Term Digital Preservation?

- *Long Term Digital Preservation (LTDP)* is a means of keeping digital information such that the same information can be used at some point in the future in spite of obsolescence of everything: hardware, software, processes, format, people, etc.
  - *Bit Preservation* addresses obsolescence of hardware
  - *Information or Logical Preservation* addresses obsolescence of everything else



Yedioth Ahronoth, Sunday June 1, 2008

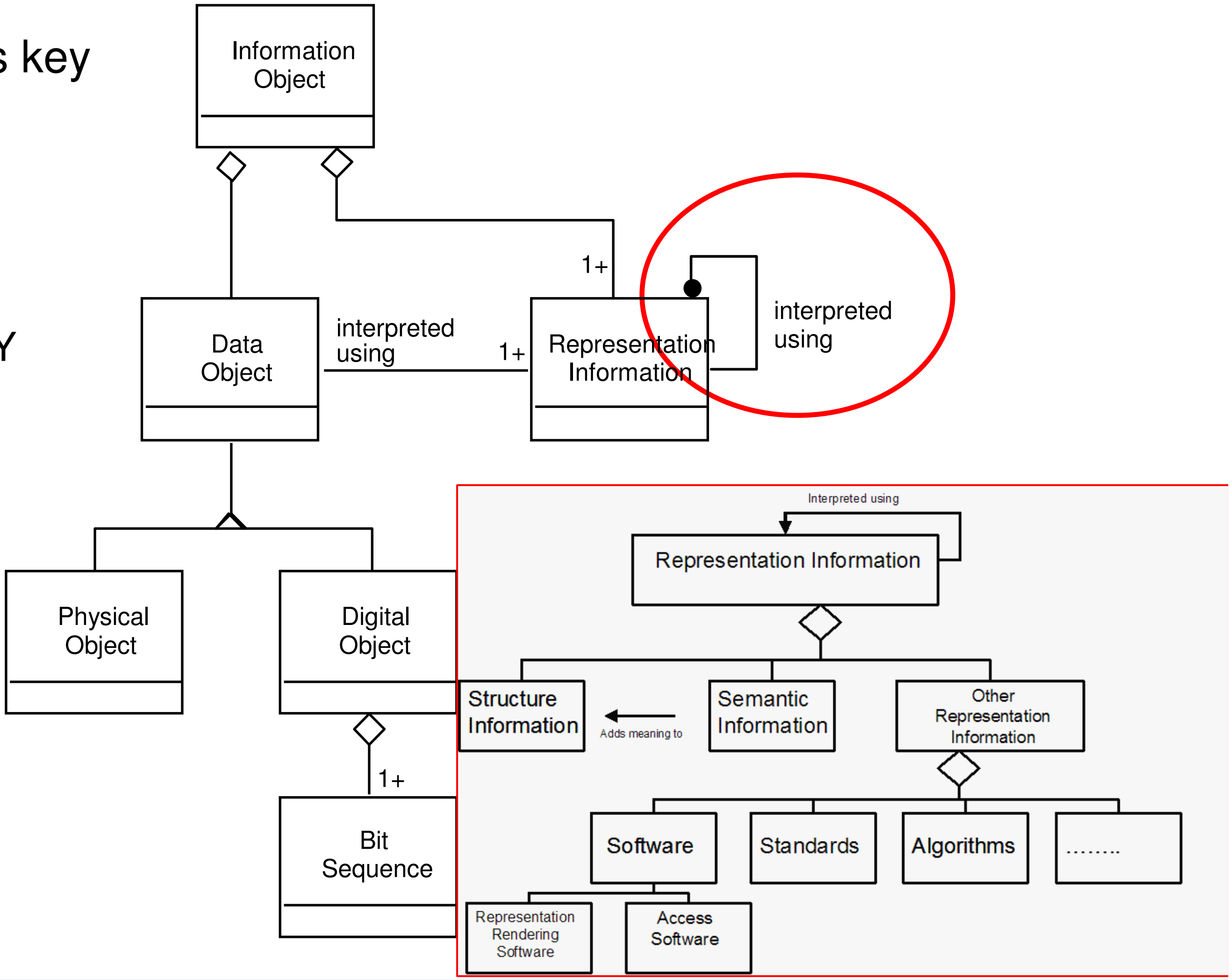
ידיעות אחרונות יום ראשון, כ"ז באייר תשס"ח - 01.06.2008

# OAIS Information Model & Representation Information

The Information Model is key

Recursion ends at KNOWLEDGEBASE of the DESIGNATED COMMUNITY

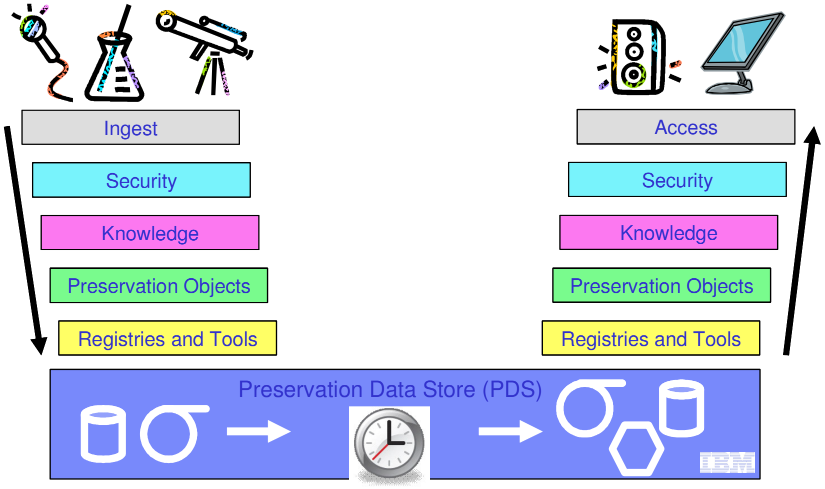
(this knowledge will change over time and region)



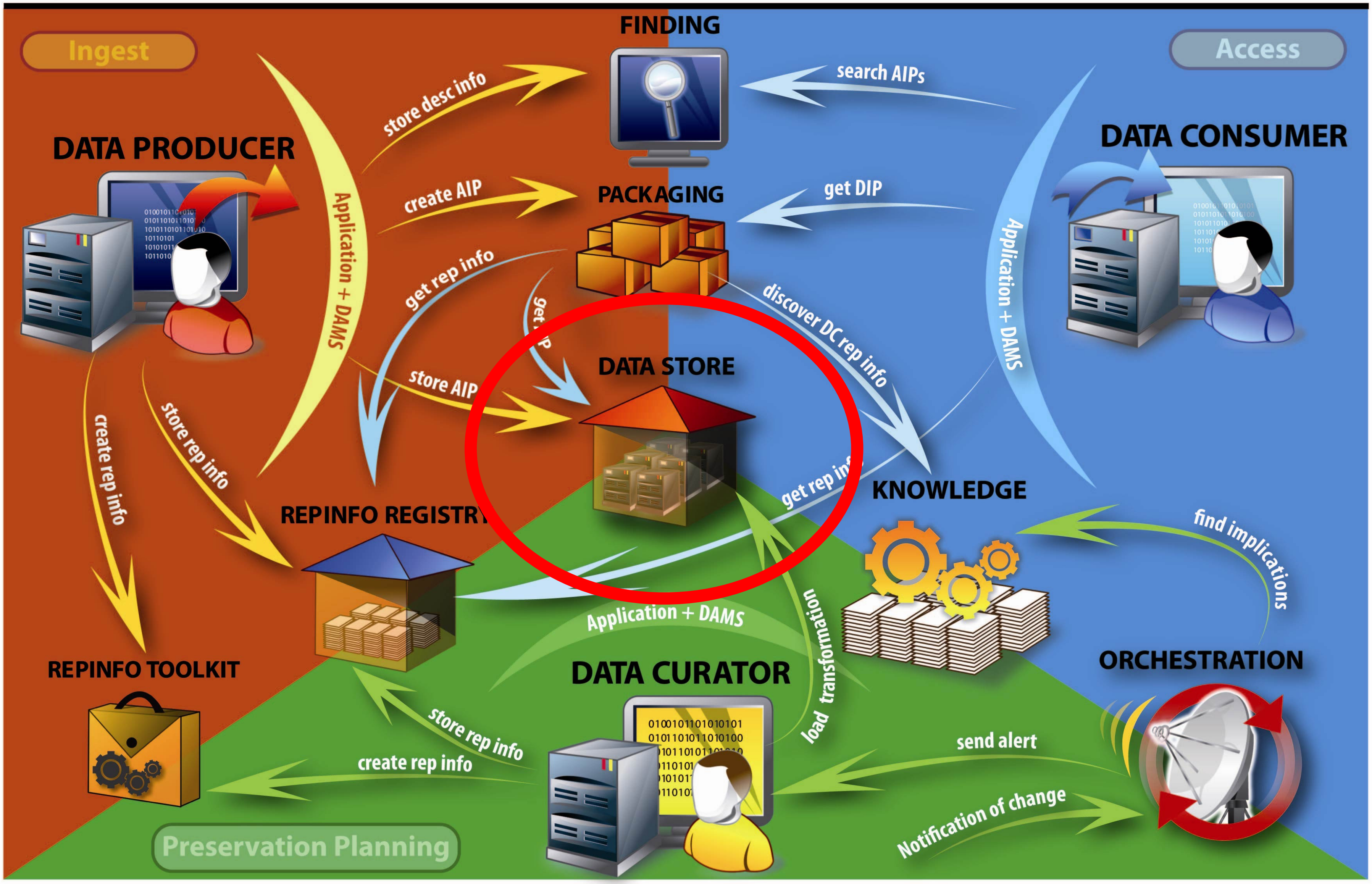
Based upon presentation by David Giaretta at EVA/MINERVA 2006: The Third Annual Jerusalem Conference on the Digitisation of Cultural Heritage

# CASPAR and Preservation DataStores

- Objective: Demonstrate validity of OAIS framework with heterogeneous data
  - Open Archival Information System (OAIS) is an ISO Standard Reference Model for Long Term Preservation
- Preservation DataStores (PDS)
  - IBM's responsibility
  - Show how to build standards-based storage that is preservation aware:
    - OAIS-required metadata
    - Transform formats to avoid information obsolescence
    - Manage media for bit obsolescence
- Other partners include data providers such as ESA, UNESCO, etc.
- **On June 24th, CASPAR (including PDS) successfully demonstrated preservation of heterogeneous data to the project officer of the European Union.**



http://www.casparpreserves.eu/ -- http://www.haifa.il.ibm.com/projects/storage/datastores/caspar.html

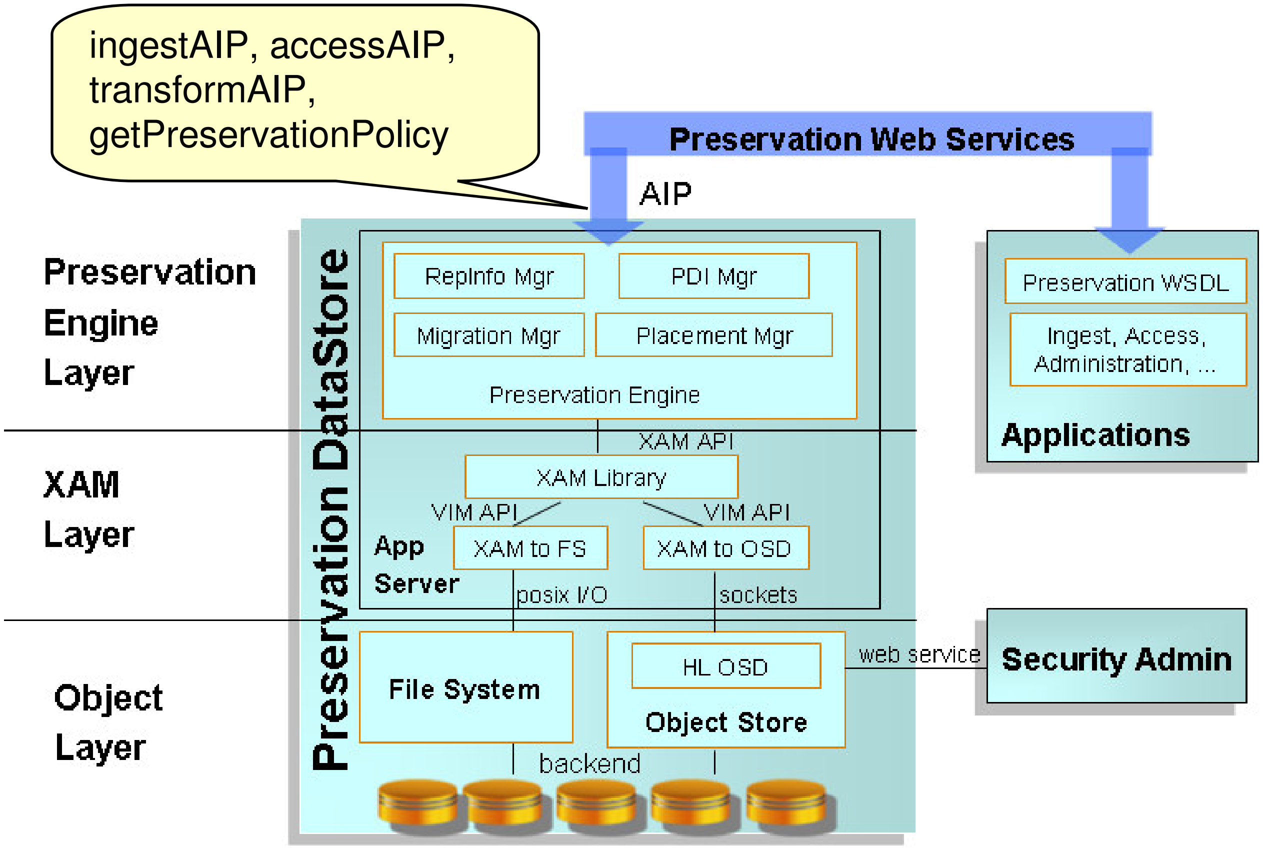

Ingest
FINDING
Access
store desc info
search AIPs
DATA PRODUCER
DATA CONSUMER
create AIP
PACKAGING
get DIP
Application + DAMS
get rep info
get AIP
discover DC rep info
Application + DAMS
DATA STORE
store AIP
store rep info
create rep info
get rep info
KNOWLEDGE
REPINFO REGISTRY
find implications
Application + DAMS
load transformation
ORCHESTRATION
REPINFO TOOLKIT
DATA CURATOR
store rep info
create rep info
send alert
Notification of change
Preservation Planning

# CASPAR Testbed Data

- Scientific data
  - European Space Agency (ESA) – IT; Science and Technology Facilities Council (STFC) – UK
  - Complex digital objects, oriented towards processing, may be high-volume
- Cultural heritage
  - UNESCO
  - Dynamic interactive digital objects, oriented towards presentation and replay
- Artistic data
  - INA, University of Leeds, CIANT, CNRS, Ircam
  - Virtual digital objects, spanning between processing and display

# PDS Architecture

- Layered approach based on open standards - OAIS, XAM, OSD
- In CASPAR, all layers are utilized. In other embodiments, only some layers can be used
- Utilize XAM to provide logical abstraction of containers (XSets)
- Offload preservation functionality to the storage



ingestAIP, accessAIP, transformAIP, getPreservationPolicy

Preservation Web Services

AIP

Preservation Engine Layer

XAM Layer

Object Layer

Preservation DataStore

RepInfo Mgr

PDI Mgr

Migration Mgr

Placement Mgr

Preservation Engine

XAM API

XAM Library

VIM API

VIM API

App Server

XAM to FS

XAM to OSD

posix I/O

sockets

File System

HL OSD

Object Store

backend

web service

Preservation WSDL

Ingest, Access, Administration, ...

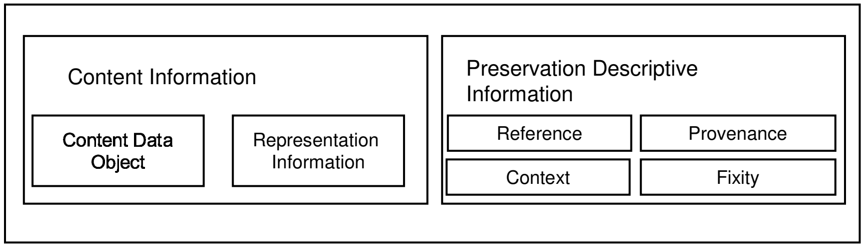
Applications

Security Admin

# Preservation DataStores (PDS)

- OAIS–based preservation-aware storage, media-agnostic and generic storage to support logical preservation

- Manage preservation specific metadata
  - Fixity computations
  - Update technical provenance
  - Manage the PDI RepInfo
  - Ensuring referential integrity



Content Information

Content Data Object

Representation Information

Preservation Descriptive Information

Reference

Provenance

Context

Fixity

- Storlet container
  - Module container that can execute restricted modules with predefined interfaces for data intensive functions, e.g., transformations, fixity calculation.
    - Optimal scheduling
  - Update PDS modules (e.g., fixity algorithm, packaging format)

- Managing availability/ data loss
  - Physically co-locate data and metadata
  - Cluster Related AIPs on the same media unit based upon their relative importance

- AIP identifier generation – Globally unique identifiers

# Preservation of an AIP

- Ingest AIP
  - Storing an AIP in PDS

- Bit and logical preservation of an AIP
  - Bit migrations
  - Format migrations - data transformation
    - Transformation modules are packed as AIPs and preserved
    - Transformation result is a new version to the original AIP
  - Migrations are documented as Provenance records
  - During migrations PDS performs operations on AIP
    - Update PDI (e.g. Fixity calculation, additional Provenance events)
    - Execute previously loaded storlets

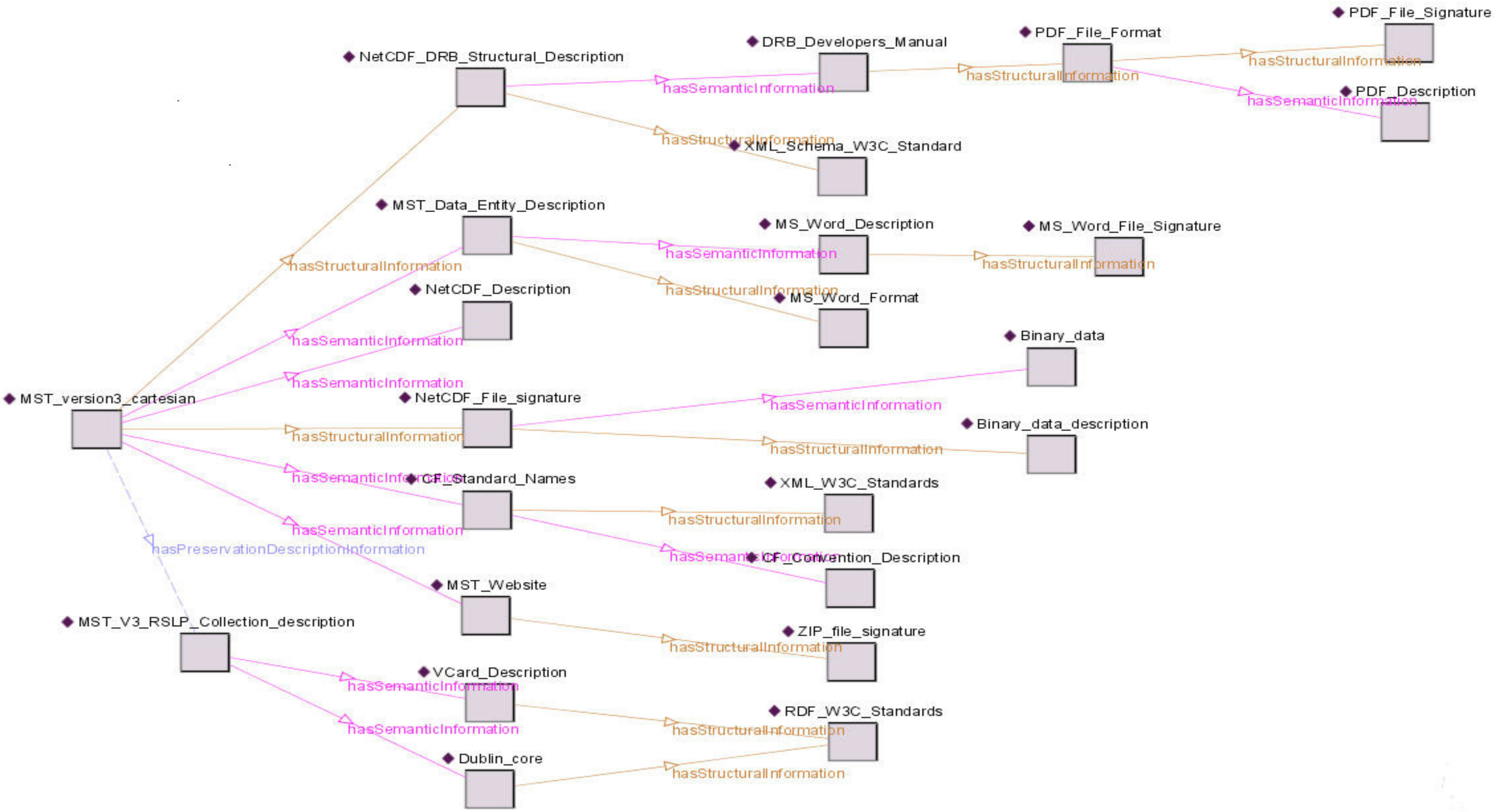
- Access AIP
  - Retrieval of an AIP
    - By retrieval time the original AIP may have several versions and copies

# MST data and PDS

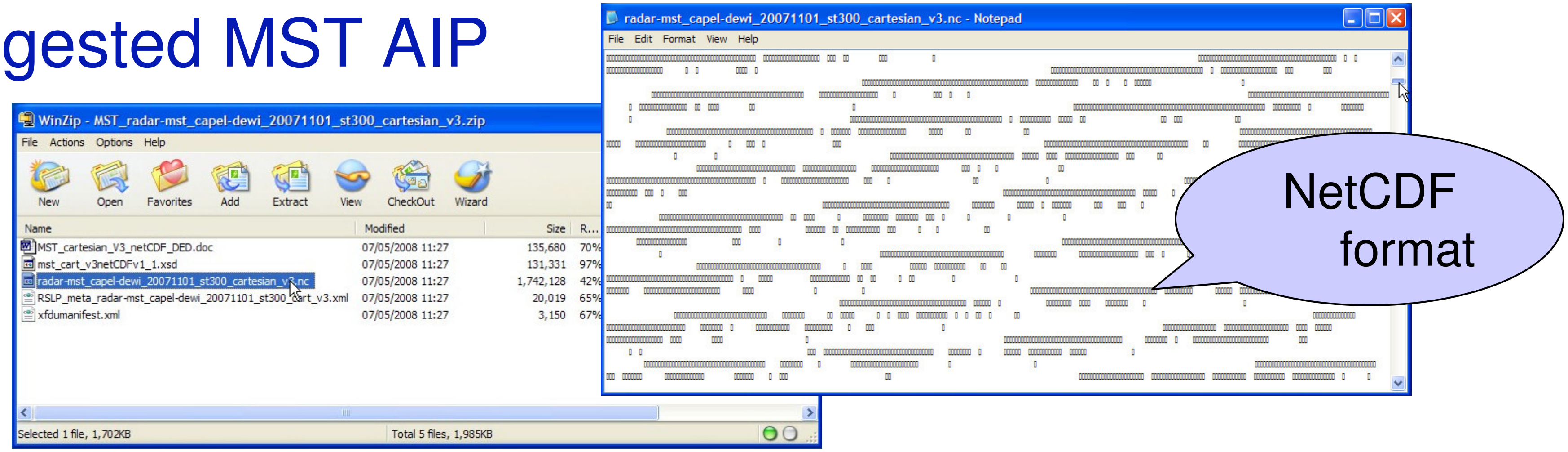
## The Natural Environment Research Council (NERC) Mesosphere-Stratosphere-Troposphere (MST) Radar at Aberystwyth

- MST data ingested to PDS
  - Highly complex atmospheric data preserved for long periods in self-describing NetCDF binary format
- PDS handles logical preservation
  - Handle metadata
    - Update provenance, compute fixity etc.
  - Load and execute transformation
    - The NASA-AMES format has become the preferred data format of the atmospheric scientists community
- Access
  - By access time there are several versions to the original AIP

# Example: MST RepInfo Network



# Ingested MST AIP



WinZip - MST_radar-mst_capel-dewi_20071101_st300_cartesian_v3.zip

radar-mst_capel-dewi_20071101_st300_cartesian_v3.nc - Notepad

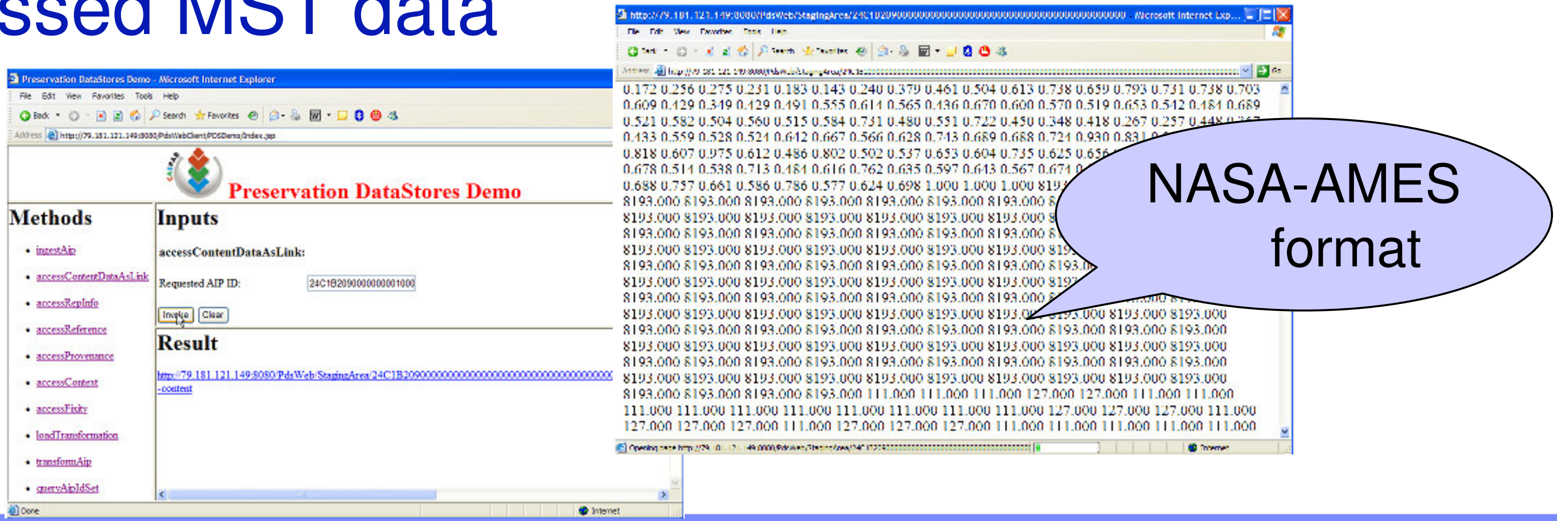
| Name | Modified | Size | R... |
|---|---|---|---|
| MST_cartesian_V3_netCDF_DED.doc | 07/05/2008 11:27 | 135,680 | 70% |
| mst_cart_v3netCDFv1_1.xsd | 07/05/2008 11:27 | 131,331 | 97% |
| radar-mst_capel-dewi_20071101_st300_cartesian_v3.nc | 07/05/2008 11:27 | 1,742,128 | 42% |
| RSLP_meta_radar-mst_capel-dewi_20071101_st300_cart_v3.xml | 07/05/2008 11:27 | 20,019 | 65% |
| xfdumanifest.xml | 07/05/2008 11:27 | 3,150 | 67% |

Selected 1 file, 1,702KB

Total 5 files, 1,985KB

NetCDF format

# Accessed MST data



Preservation DataStores Demo - Microsoft Internet Explorer

http://79.181.121.149:8080/PdsWebClient/PDSDemo/index.jsp

Preservation DataStores Demo

**Methods**

- ingestAip
- accessContentDataAsLink
- accessRepInfo
- accessReference
- accessProvenance
- accessContext
- accessFixity
- loadTransformation
- transformAip
- queryAipIdSet

**Inputs**

accessContentDataAsLink:

Requested AIP ID: 24C1B209000000001000

Invoke Clear

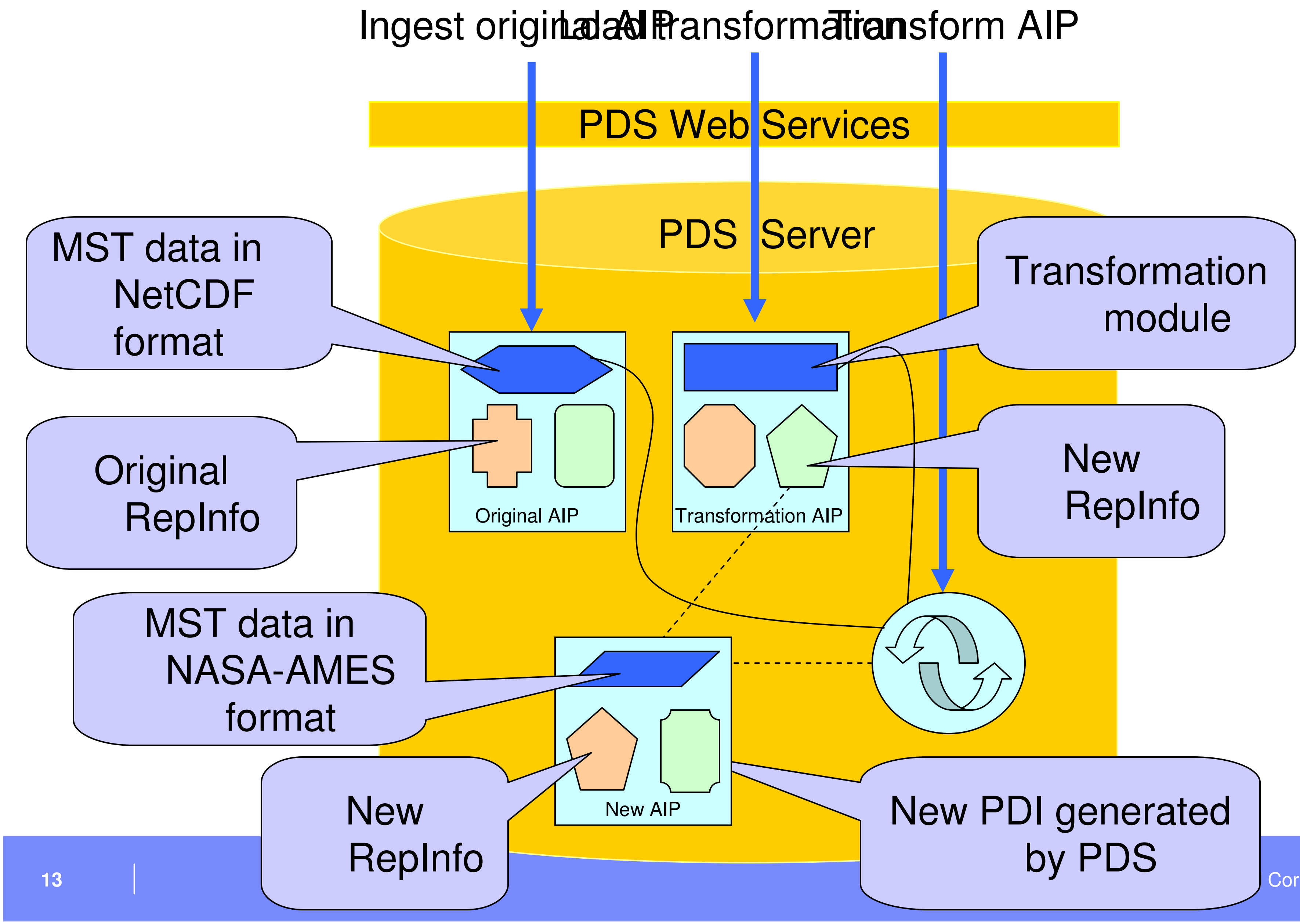
**Result**

http://79.181.121.149:8080/PdsWeb/StagingArea/24C1B2090000000000000000000000000000000000-content

http://79.181.121.149:8080/PdsWeb/StagingArea/24C1B2090000000000000000000000000000000000000000000 - Microsoft Internet Exp...

```
0.172 0.256 0.275 0.231 0.183 0.143 0.240 0.379 0.461 0.504 0.613 0.738 0.659 0.793 0.731 0.738 0.703
0.609 0.429 0.349 0.429 0.491 0.555 0.614 0.565 0.436 0.670 0.600 0.570 0.519 0.653 0.542 0.484 0.689
0.521 0.582 0.504 0.560 0.515 0.584 0.731 0.480 0.551 0.722 0.450 0.348 0.418 0.267 0.257 0.448
0.433 0.559 0.528 0.524 0.642 0.667 0.566 0.628 0.743 0.689 0.688 0.724 0.930 0.831
0.818 0.607 0.975 0.612 0.486 0.802 0.502 0.537 0.653 0.604 0.735 0.625 0.656
0.678 0.514 0.538 0.713 0.484 0.616 0.762 0.635 0.597 0.643 0.567 0.674
0.688 0.757 0.661 0.586 0.786 0.577 0.624 0.698 1.000 1.000 1.000
8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000
8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000
8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000
8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000
8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000
8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000
8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000
8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000
8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000
8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000
8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000
8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000
8193.000 8193.000 8193.000 8193.000 111.000 111.000 111.000 127.000 127.000 111.000 111.000
111.000 111.000 111.000 111.000 111.000 111.000 111.000 111.000 127.000 127.000 127.000 111.000
127.000 127.000 127.000 111.000 127.000 127.000 127.000 111.000 111.000 111.000 111.000 111.000
```

NASA-AMES format



Ingest original AIP

Load AIP Transformation

Transform AIP

PDS Web Services

PDS Server

MST data in NetCDF format

Original RepInfo

MST data in NASA-AMES format

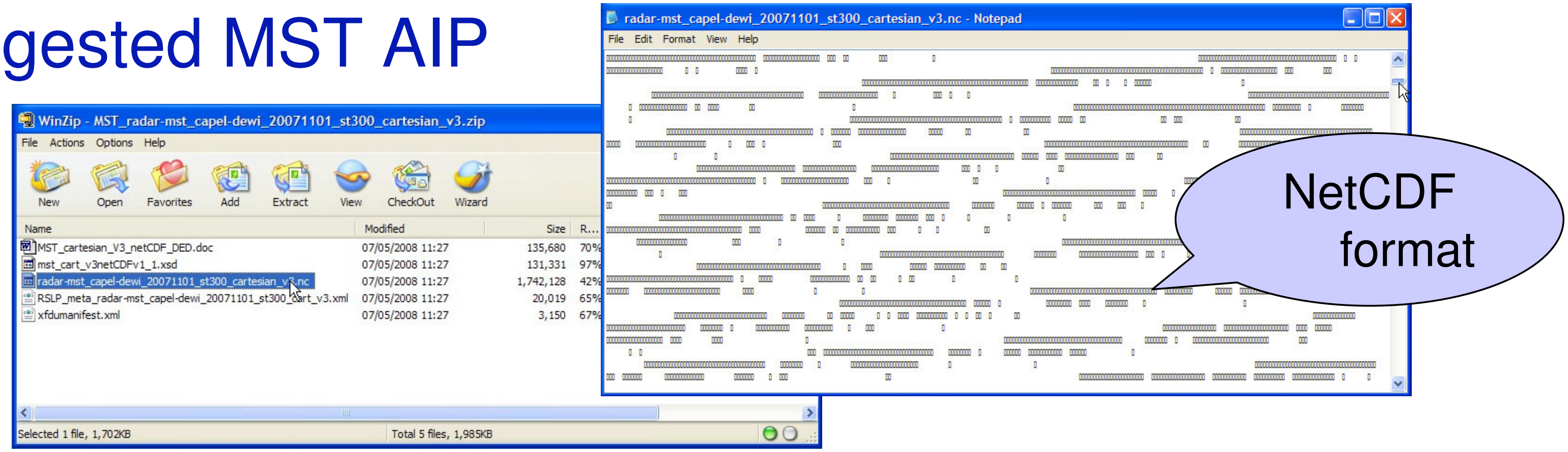
New RepInfo

Transformation module

New RepInfo

New PDI generated by PDS

# Ingested MST AIP



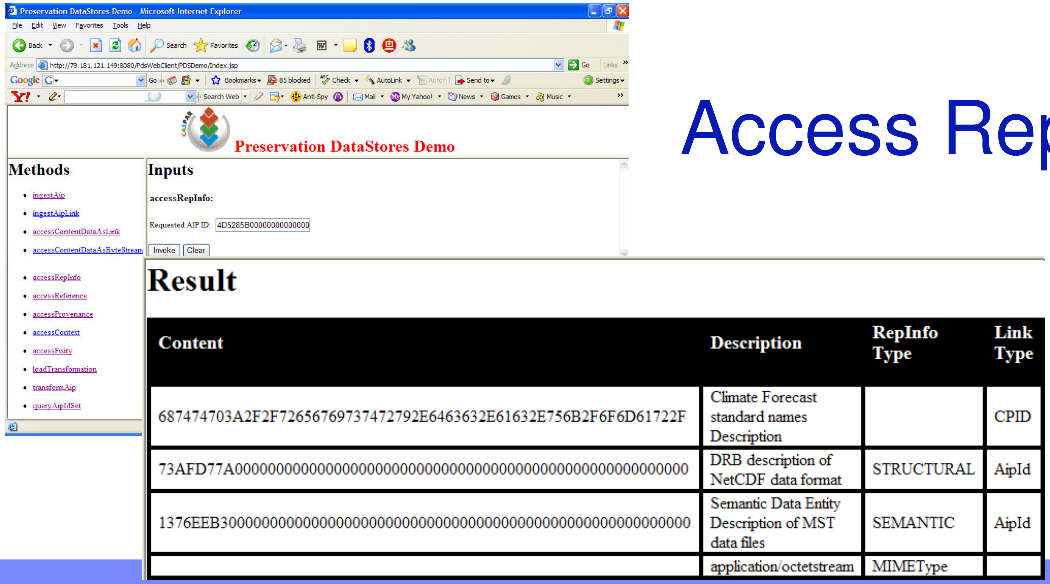
WinZip - MST_radar-mst_capel-dewi_20071101_st300_cartesian_v3.zip
File Actions Options Help
New Open Favorites Add Extract View CheckOut Wizard
radar-mst_capel-dewi_20071101_st300_cartesian_v3.nc - Notepad
File Edit Format View Help
NetCDF format

| Name | Modified | Size | R... |
|---|---|---|---|
| MST_cartesian_V3_netCDF_DED.doc | 07/05/2008 11:27 | 135,680 | 70% |
| mst_cart_v3netCDFv1_1.xsd | 07/05/2008 11:27 | 131,331 | 97% |
| radar-mst_capel-dewi_20071101_st300_cartesian_v3.nc | 07/05/2008 11:27 | 1,742,128 | 42% |
| RSLP_meta_radar-mst_capel-dewi_20071101_st300_cart_v3.xml | 07/05/2008 11:27 | 20,019 | 65% |
| xfdumanifest.xml | 07/05/2008 11:27 | 3,150 | 67% |

Selected 1 file, 1,702KB

Total 5 files, 1,985KB

# Access RepInfo of original AIP



Preservation DataStores Demo - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://79.181.121.149:8080/PdsWebClient/PDSDemo/Index.jsp
Go Links
Google Go Bookmarks 85 blocked Check AutoLink AutoFill Send to Settings
Search Web Anti-Spy Mail My Yahoo! News Games Music
CASPAR
Invoke Clear

Preservation DataStores Demo

**Methods**

- ingestAip
- ingestAipLink
- accessContentDataAsLink
- accessContentDataAsByteStream

- accessRepInfo
- accessReference
- accessProvenance
- accessContext
- accessFixity
- loadTransformation
- transformAip
- queryAipIdSet

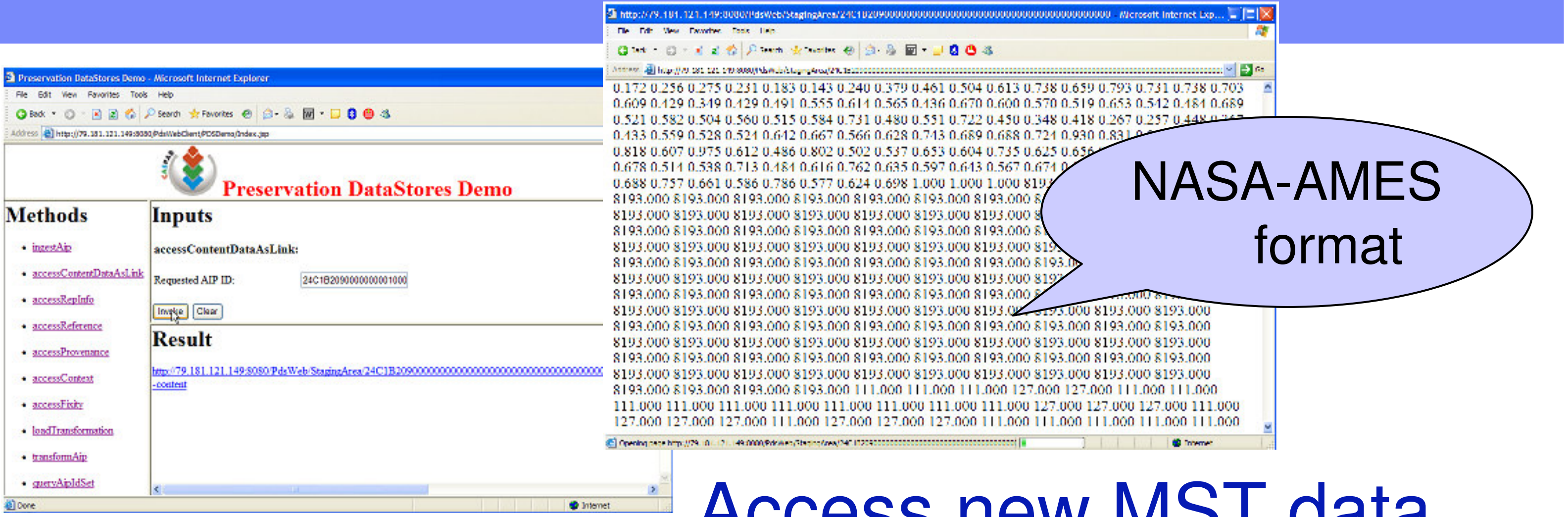
**Inputs**

**accessRepInfo:**

Requested AIP ID: 4D5285B0000000000000

## Result

| Content | Description | RepInfo Type | Link Type |
|---|---|---|---|
| 687474703A2F2F72656769737472792E6463632E61632E756B2F6F6D61722F | Climate Forecast standard names Description | | CPID |
| 73AFD77A00000000000000000000000000000000000000000000000000000000 | DRB description of NetCDF data format | STRUCTURAL | AipId |
| 1376EEB300000000000000000000000000000000000000000000000000000000 | Semantic Data Entity Description of MST data files | SEMANTIC | AipId |
| | application/octetstream | MIMEType | |



Preservation DataStores Demo - Microsoft Internet Explorer

Address: http://79.181.121.149:8080/PdsWebClient/PDSDemo/Index.jsp

Preservation DataStores Demo

**Methods**

- ingestAip
- accessContentDataAsLink
- accessRepInfo
- accessReference
- accessProvenance
- accessContext
- accessFixity
- loadTransformation
- transformAip
- queryAipIdSet

**Inputs**

accessContentDataAsLink:

Requested AIP ID: 24C1B209000000001000

Invoke Clear

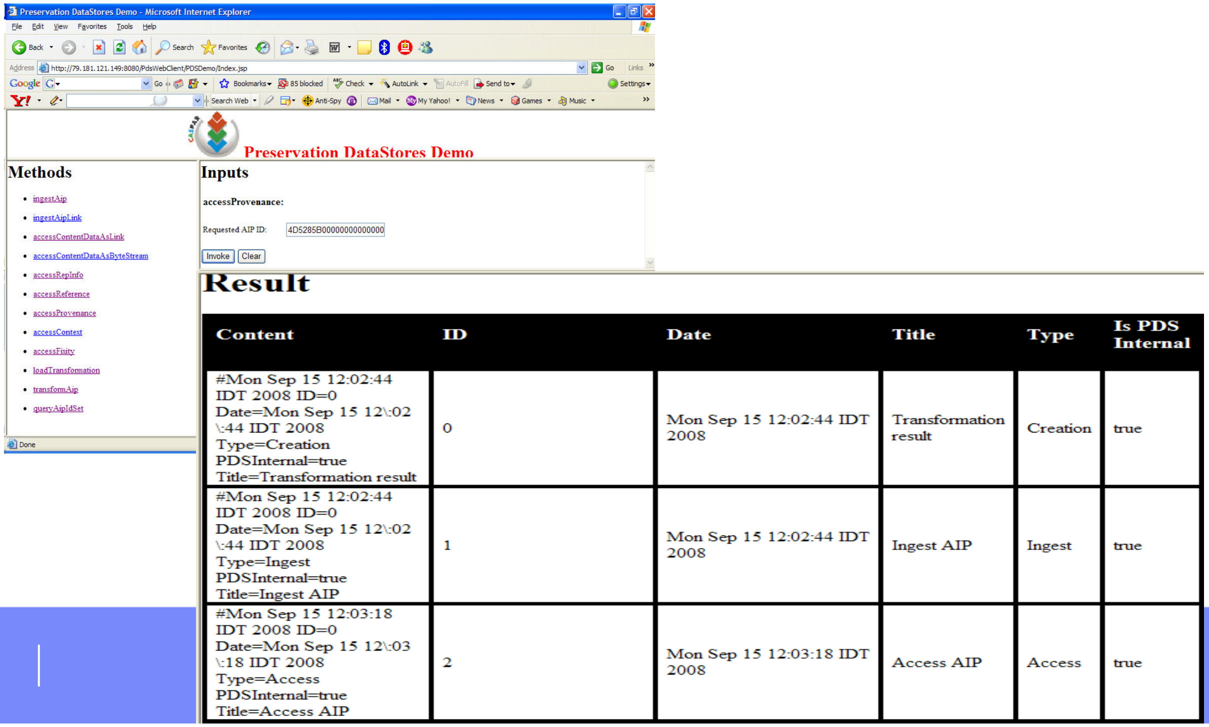
**Result**

http://79.181.121.149:8080/PdsWeb/StagingArea/24C1B2090000000000000000000000000000000
-content

http://79.181.121.149:8080/PdsWeb/StagingArea/24C1B209000000000000000000000000000000000000000 - Microsoft Internet Exp...

```
0.172 0.256 0.275 0.231 0.183 0.143 0.240 0.379 0.461 0.504 0.613 0.738 0.659 0.793 0.731 0.738 0.703
0.609 0.429 0.349 0.429 0.491 0.555 0.614 0.565 0.436 0.670 0.600 0.570 0.519 0.653 0.542 0.484 0.689
0.521 0.582 0.504 0.560 0.515 0.584 0.731 0.480 0.551 0.722 0.450 0.348 0.418 0.267 0.257 0.448
0.433 0.559 0.528 0.524 0.642 0.667 0.566 0.628 0.743 0.689 0.688 0.721 0.930 0.831
0.818 0.607 0.975 0.612 0.486 0.802 0.502 0.537 0.653 0.604 0.735 0.625 0.656
0.678 0.514 0.538 0.713 0.484 0.616 0.762 0.635 0.597 0.643 0.567 0.674
0.688 0.757 0.661 0.586 0.786 0.577 0.624 0.698 1.000 1.000 1.000
8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000
8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000
8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000
8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000
8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000
8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000
8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000
8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000
8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000
8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000
8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000
8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000 8193.000
8193.000 8193.000 8193.000 8193.000 111.000 111.000 111.000 127.000 127.000 111.000 111.000
111.000 111.000 111.000 111.000 111.000 111.000 111.000 111.000 127.000 127.000 127.000 111.000
127.000 127.000 127.000 111.000 127.000 127.000 127.000 111.000 111.000 111.000 111.000 111.000
```

## Access new MST data



Preservation DataStores Demo - Microsoft Internet Explorer

Address: http://79.181.121.149:8080/PdsWebClient/PDSDemo/Index.jsp

Preservation DataStores Demo

**Methods**

- ingestAip
- ingestAipLink
- accessContentDataAsLink
- accessContentDataAsByteStream
- accessRepInfo
- accessReference
- accessProvenance
- accessContext
- accessFixity
- loadTransformation
- transformAip
- queryAipIdSet

**Inputs**

accessProvenance:

Requested AIP ID: 4D5285B00000000000000

Invoke Clear

**Result**

| Content | ID | Date | Title | Type | Is PDS Internal |
|---|---|---|---|---|---|
| #Mon Sep 15 12:02:44 IDT 2008 ID=0 Date=Mon Sep 15 12\:02 \:44 IDT 2008 Type=Creation PDSInternal=true Title=Transformation result | 0 | Mon Sep 15 12:02:44 IDT 2008 | Transformation result | Creation | true |
| #Mon Sep 15 12:02:44 IDT 2008 ID=0 Date=Mon Sep 15 12\:02 \:44 IDT 2008 Type=Ingest PDSInternal=true Title=Ingest AIP | 1 | Mon Sep 15 12:02:44 IDT 2008 | Ingest AIP | Ingest | true |
| #Mon Sep 15 12:03:18 IDT 2008 ID=0 Date=Mon Sep 15 12\:03 \:18 IDT 2008 Type=Access PDSInternal=true Title=Access AIP | 2 | Mon Sep 15 12:03:18 IDT 2008 | Access AIP | Access | true |

## Access new AIP provenance

# Some other things we are doing in IBM Research

- Storage Networking Industry Association (SNIA)
  - LTACSI - Long Term Archive & Compliance Storage Initiative
  - IBM co-chairs the Long Term Retention TWG
    - A recently-formed working group focusing on a Self-contained Information Retention Format
- Long Term Digital Preservation Assessment
  - Research tool to evaluate organization's ability to preserve its digital resources
  - Based upon emerging standard audit checklists (ISO 14721)

# More Info

http://www.haifa.il.ibm.com/projects/storage/ltdp/index.shtml

Michael Factor: factor@il.ibm.com

IBM Haifa Research Lab