SHAMAN

Sustaining Heritage Access through Multivalent ArchiviNg

SHAMAN and Storage Virtualization

Adil Hasan
Univ. of Liverpool

Toulouse Workshop on Storage Virtualization
2nd June 2009

SEVENTH FRAMEWORK PROGRAMME

Co-funded by the European Union

inmark
ESTUDIOS Y ESTRATEGIAS, S. A.

UNIVERSITY OF LIVERPOOL

In|Con|Tec

HÖGSKOLAN I BORÅS

XEROX®

FernUniversität in Hagen

PHILIPS

CENTRE FOR DIGITAL LIBRARY RESEARCH
cdlr.strath.ac.uk

DEUTSCHE NATIONAL BIBLIOTHEK

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

OTTO-VON-GUERICKE-UNIVERSITÄT MAGDEBURG

Industrious Media

GLOBIT

University of Glasgow | Humanities Advanced Technology & Information Institute

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

inesc id lisboa

SHAMAN Collaborators:

SDSC

VLAAMS PARLEMENT

SHAMAN

Sustaining Heritage Access through Multivalent ArchiviNg

What is SHAMAN?

- Sustaining HeritAge through Multivalent ArchiviNg.
- FP7 EU Integrated Project start Dec/07 finish Dec/11.
- Aim to investigate long-term preservation of data-sets.

What is SHAMAN?

- Issues:
  - Accommodate unknown changes to hardware (infrastructure).
  - Accommodate unknown changes to preservation tools (processes).
  - Accommodate unknown changes to format and description of data (content).

Infrastructure

- Hardware will change:
  - Provide infrastructure layer on-top of hardware.
  - Encapsulate changes, only hardware layer worries about changes.
  - Uniform interface to hardware; driver maps from hardware interface to infrastructure interface.
  - Logical-to-physical object mapping; insulates from changes in data location.
  - Record characteristics of hardware (for debugging).

Preservation Process

- Components of process will change:
  - Use abstract language to describe preservation processes (these are the policies).
  - Translate from abstract language to actual work-flows.
  - Insulates from changes to work-flow engines or services.
  - Abstract description of processes must be preserved.

Content

- Format and description of data will change:
  - Try to keep data in original format (save space), render to user with migratable tool.
  - Migrate data-sets that cannot be rendered reliably (closed source format, complex, etc).
  - Use standards for description of data.
  - Ensure dictionary of terminology (domain-specific terms) archived.

PDI

- Preservation of a digital object requires:
  - data-format, data-description, processes, hardware
- Can regard this collection as the context of the data.
- Want to capture this context in a single unit; the Preservation Description Information.
- PDI must be related to data.
- PDI itself must be preserved.

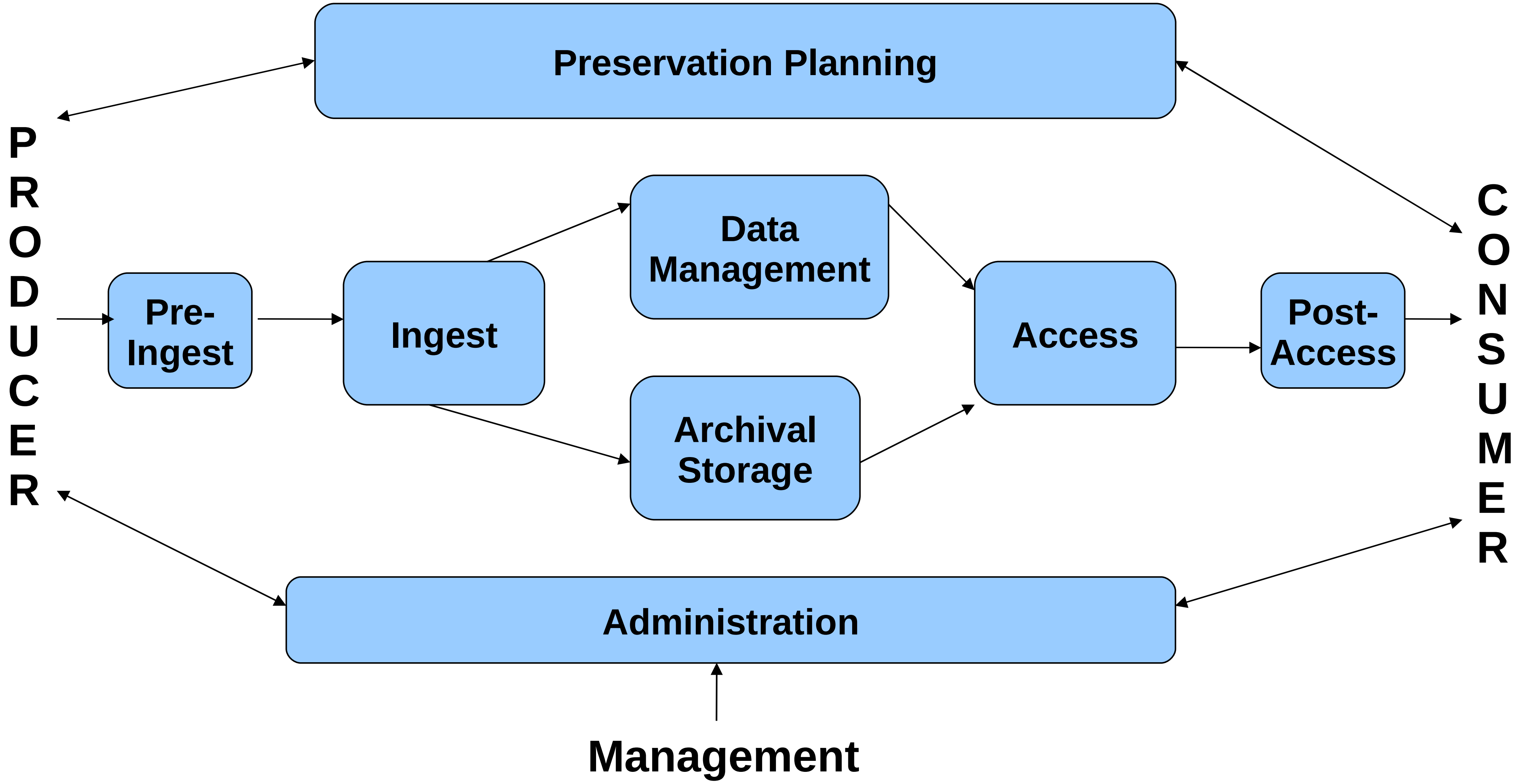
extended-OAIS

- OAIS describes long-term archive system.
- Used in many projects. Is the basis for SHAMAN.
- Needs improvements:
  - Preserving process information necessary.
  - Pre-Ingest phase required (recognise importance of assembling data for ingest).
  - Post-Access phase required (recognise importance of further processing after extraction of data).

SHAMAN

Sustaining Heritage Access through Multivalent ArchiviNg

extended-OAIS



extended-OAIS

- Further observations:
  - Preservation Planning must encompass Ingest and Access.
  - Preservation Planning requires input from Producer, Consumer and Management.
  - Roles must be further refined to understand better mapping to existing business roles.

Storage

- Essentially 3-types of storage:
  - Ingest Storage – holds ingested data where AIP is created. Temporary Storage
  - Dissemination Storage – holds data for access where DIP is assembled. Temporary Storage.
  - Archival Storage – holds archived data and description (i.e. AIP). Permanent Storage.

Storage

- Ingest and Dissemination are most-likely frequent-access systems.
    - Viewed as cache-systems for the Archival Storage.
    - System must have good performance.
- Archival storage less-frequently access, must be reliable.

Storage

- Amount of digital data increasing rapidly.
- Many institutions collaborate.
- Makes sense to make use of collaborators storage.
- Allows storage to scale with data volume.
- Geographically distributed data guards against storage failure.

SHAMAN
Sustaining Heritage Access through Multivalent ArchiviNg

Data Grids

- Data Grids provide a means of combining distributed resources into logical resource.
- Insulate storage from access to storage.
  - Provide uniform access to resources through drivers.
- Removes dependence on physical location through logical-physical file mapping.
- Virtualization of storage.

Data Grids

- Observations:
  - Data Grid resource may not be part of the collaboration. May be third-party.
  - Need SLA on provision of resource such that it is possible to replace resource with different type (e.g. cloud).
  - Need to ensure SLA is implemented (evidence).
  - Need to make sure preservation processes are sufficiently encapsulated to run on collaborator/external resources.

iRODs

- Data Grid provides distributed storage.
- Need to manage that storage. To apply policies to that storage.
- IRODS – policy-driven data management system.
- Allows virtualization of storage AND virtualization of policies.
- Rule-engine manages the policies.
- Policies implemented as rules.

iRODS

- Can view iRODS rule-engine as managing the preservation processes:
  - Replicate data, checksum data, transform data, extract metadata etc.
- Require tool to transform from abstract policies to rules (in preparation).
- Rule execution must be logged (which rule, microservice, result, when run) for auditing.
- Validation of each microservice for auditing.

What about the Cloud?

- Cloud, such as Amazon S3 can be viewed as another storage resource within the data grid.
- SLA would ensure Cloud resources provide necessary functionality.
  - Would need to ensure cloud providers are capable of providing mechanism to ensure SLA enforced.
  - Checks can be implemented as rules in iRODS.

Access

- Long-term access can be achieved by:
- Migration to new formats
  - Requires more storage to hold original and migrated copy. Intensive process.
  - May be only option for some closed formats.
- Rendering of old format
  - Tool capable of reading old format means 'migration' on demand.
  - Need to ensure all the required properties of data are accessible.

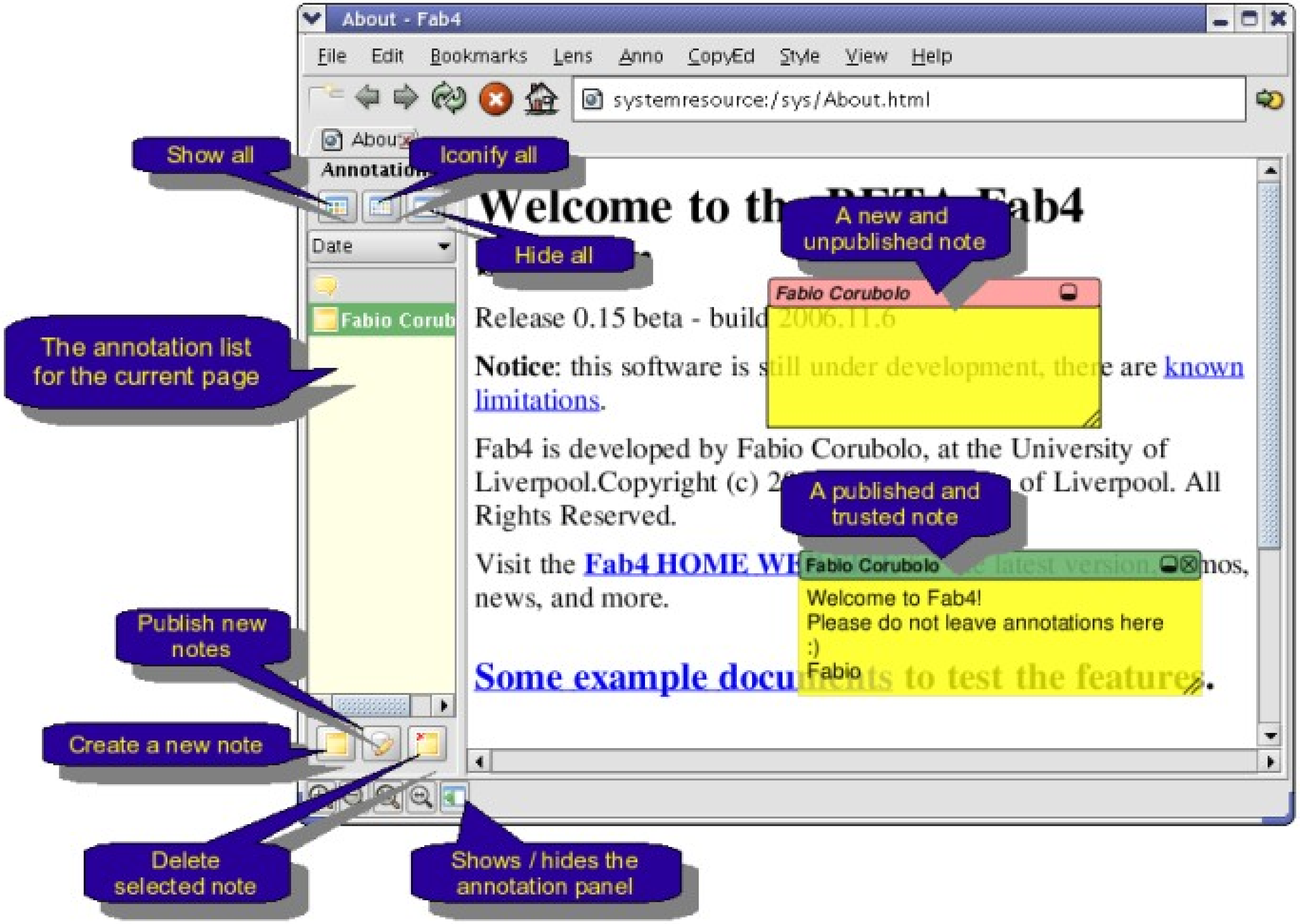
Access

- SHAMAN make use of Multivalent tool:
  - Renders data in old format to user.
  - Written in Java.
  - Capable of reading different formats through different drivers ('media engines').
- Plan to archive tool and package-up tool and data in original format for download to allow access to data.

SHAMAN

Sustaining Heritage Access through Multivalent ArchiviNg

Access


About - Fab4
File Edit Bookmarks Lens Anno CopyEd Style View Help
systemresource:/sys/About.html
About
Annotation
Date
Fabio Corub
Show all
Iconify all
Hide all
A new and unpublished note
A published and trusted note
The annotation list for the current page
Publish new notes
Create a new note
Delete selected note
Shows / hides the annotation panel
Fabio Corubolo
Fabio Corubolo

Welcome to th BETA Fab4

Release 0.15 beta - build 2006.11.6

Notice: this software is still under development, there are known limitations.

Fab4 is developed by Fabio Corubolo, at the University of Liverpool.Copyright (c) 2 of Liverpool. All Rights Reserved.

Visit the Fab4 HOME WE latest version, mos, news, and more.

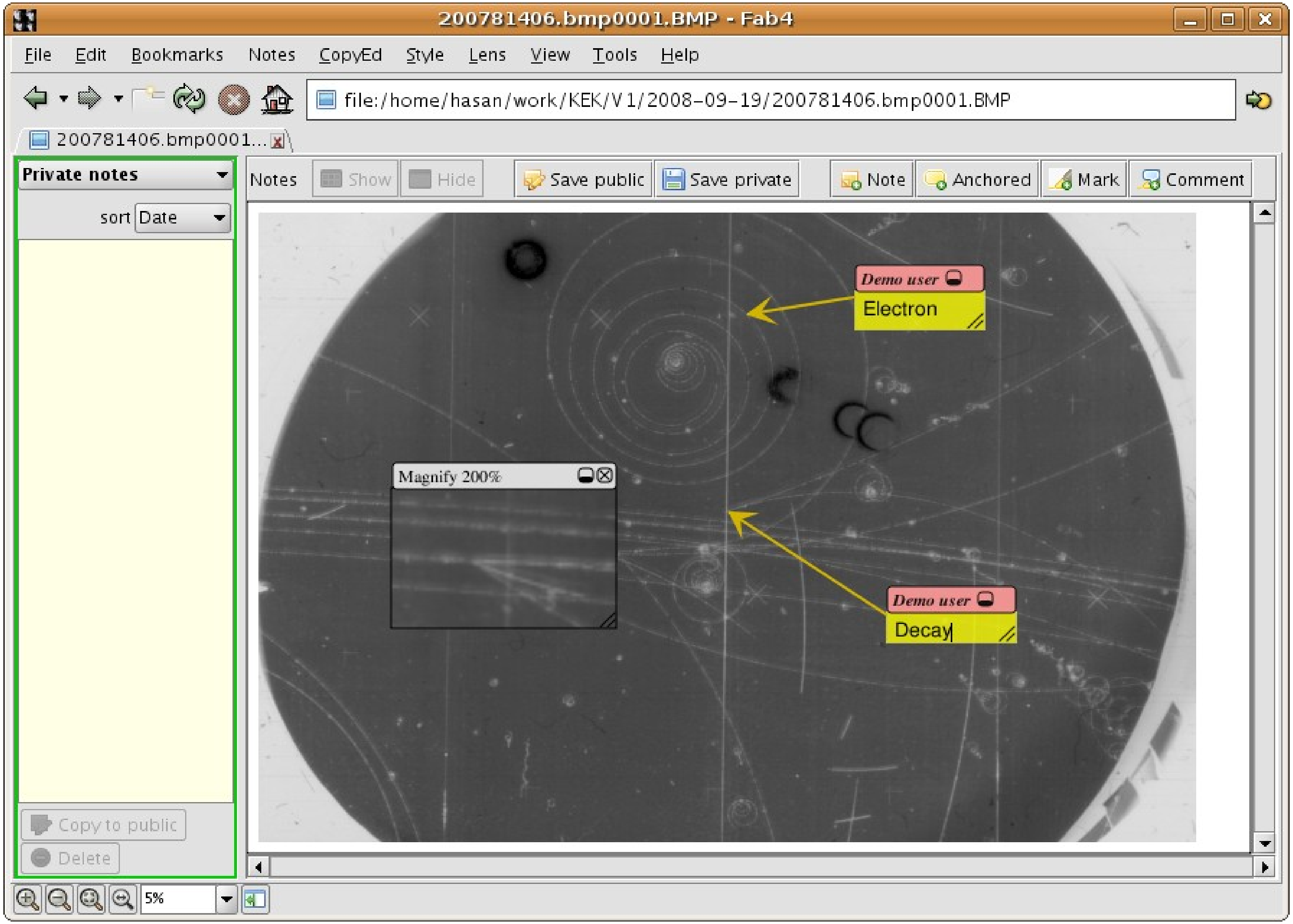
Some example documents to test the features.

Welcome to Fab4!
Please do not leave annotations here
:)
Fabio

Access

- Fab4 browser interface to Multivalent.
- Allows annotations to be made on document.
- Annotations saved in separate file from document.
- Data not altered.
- Also apply different behaviours such as magnifying lens

Access


200781406.bmp0001.BMP - Fab4
File Edit Bookmarks Notes CopyEd Style Lens View Tools Help
file:/home/hasan/work/KEK/V1/2008-09-19/200781406.bmp0001.BMP
200781406.bmp0001...
Private notes
sort Date
Notes Show Hide Save public Save private Note Anchored Mark Comment
Demo user
Electron
Magnify 200%
Demo user
Decay
Copy to public
Delete
5%

Access

- Annotations semantically attached to document.
  - For images will need to be re-worked.
- Allows annotations and behavours to appear in correct position in document in different formats.

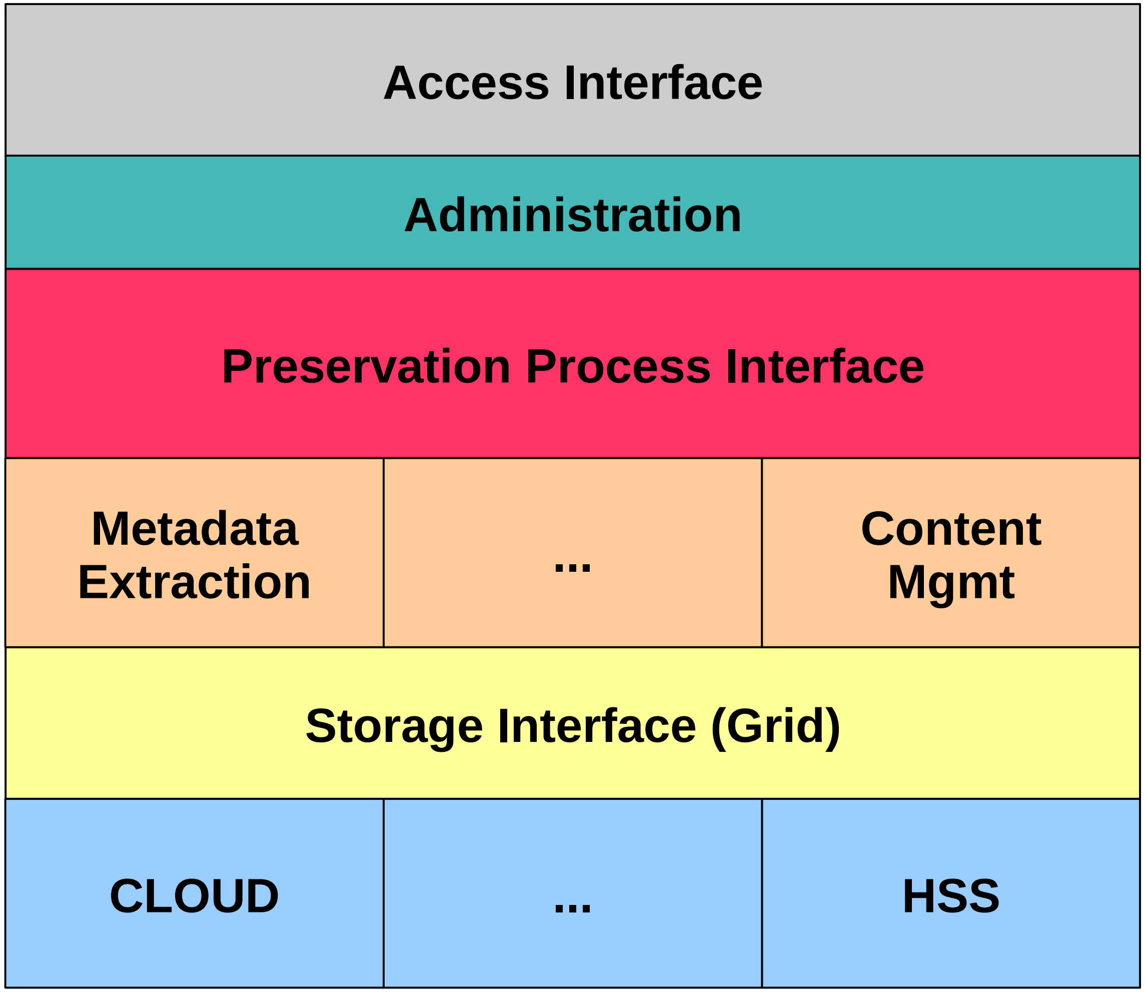
Access

- Successful discovery requires as much information about the digital object as possible is supplied during ingest.
- Also requires information is extracted and indexed.
- SHAMAN make use of powerful Cheshire digital library system to extract information such that it is discoverable.
- Processes involved in extraction must be preserved.

SHAMAN

Sustaining Heritage Access through Multivalent ArchiviNg

Preservation Layers



Access Interface

Administration

Preservation Process Interface

Metadata Extraction | ... | Content Mgmt

Storage Interface (Grid)

CLOUD | ... | HSS

Provides uniform access to data.

Manages access & system, Transforms pres processes from abstract

Interface provides uniform access to different preservation processes.

Grid interfaces to different Types of storage. Provides Uniform Interface

Summary

- To ensure data usable in the long-term:
  - Insulate from hardware changes.
  - Insulate from changes to processes.
  - Insulate from changes to data format.
  - Insulate from changes to description.
  - Ensure as much information as possible about data is captured.
    - Ideally test data is understandable without ANY external dependencies.
- SHAMAN aims to provide a framework that accounts for these issues.