Using Data Grids for Long Term Preservation (The SHAMAN Project)

Adil Hasan

University of Liverpool

What is SHAMAN?

- Sustaining HeritAge through Multivalent ArchiviNg.
- FP7 EU Integrated Project started Dec/07 finish Dec/11.
- 17 partners: DICE group (US), DNB (D), FUH (D), GLOBIT (D), HATII (GB), INCONTEC (D), INESC-ID (P), INMARK (ESP), IM (GB), Philips (NL), SSLIS (S), UGottingen (D), UIUC (US), UMagdeburg (D), ULiverpool (GB), UStrathclyde (GB), Xerox (F).

What is SHAMAN?

- Aim to investigate long-term preservation of large data-sets.
- Framework must guarantee future accessibility of data even when h/w and s/w change.
- To ensure data understandable in future must also preserve enough context information.

The SHAMAN Approach

- Decouple preservation processes from data.
  - Use abstract language to define preservation processes.
  - Possible to replace underlying services as they become obsolete.
  - Preservation processes themselves must be preserved.
- Must also preserve enough contextual information to make sense of the data in the future.

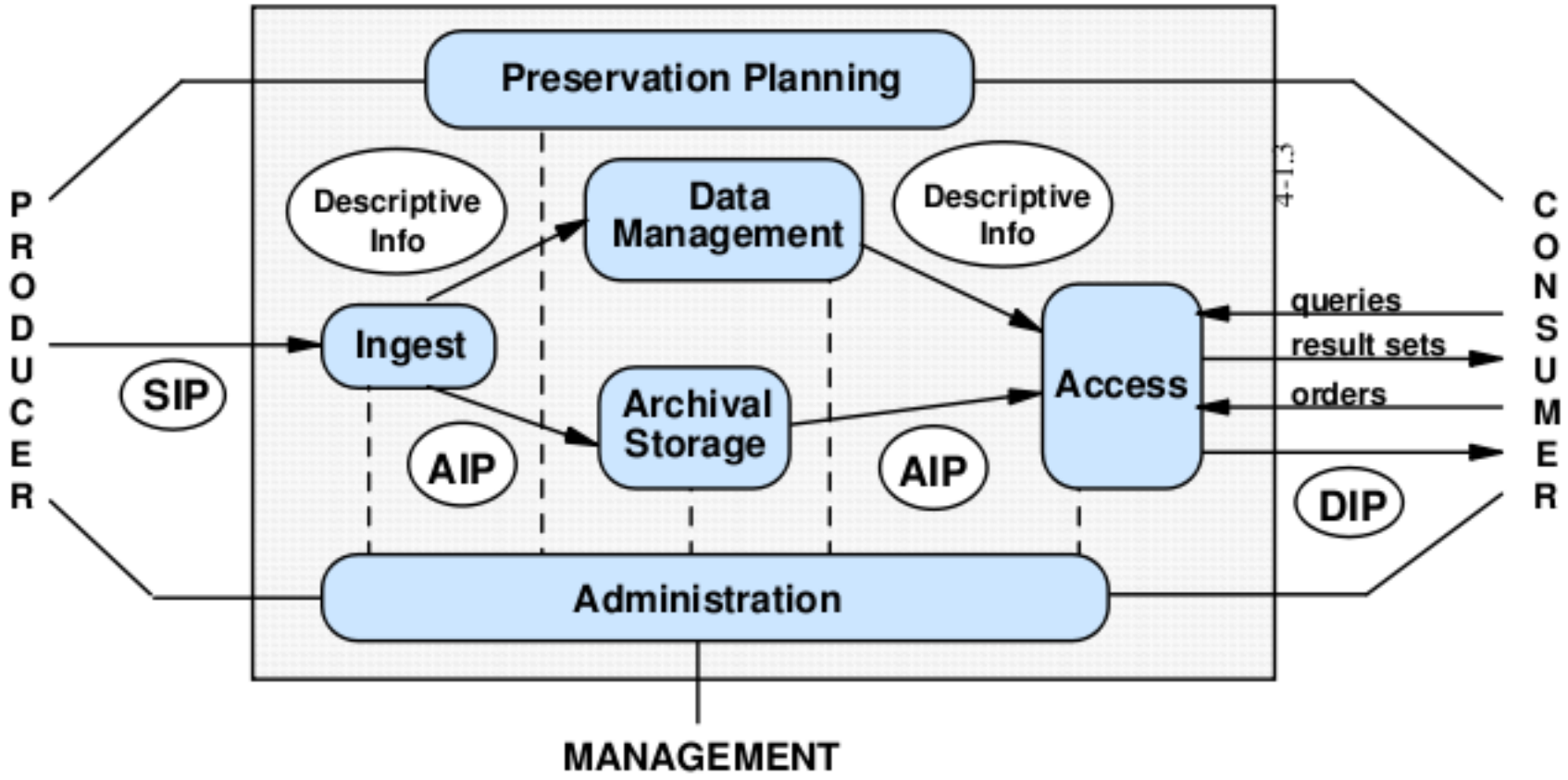
The SHAMAN Approach

- Decouple storage from data
  - Use data-grid to insulate from changes to hardware.
  - Allows system to scale by easily accommodating new hardware.
  - Allows system to interoperate with other systems through federation.

The SHAMAN Approach

- If possible, keep data in original format and use migrateable tool to render data to end-user.
  - Avoids need to migrate all data regardless of access.
  - CPU only used to render data that needs to be accessed.
  - Tool has adapters to read obsolete formats.
  - Only need to migrate the tool forwards.

Open Archival Information System



iRODS

- Provide storage virtualization
  - Use logical names for storage can replace storage.
- Provides policy virtualization
  - Can use rules to implement some preservation policies which are executed by micro-services.
- Provides a trusted archive
  - Can implement rules to check validity of data.

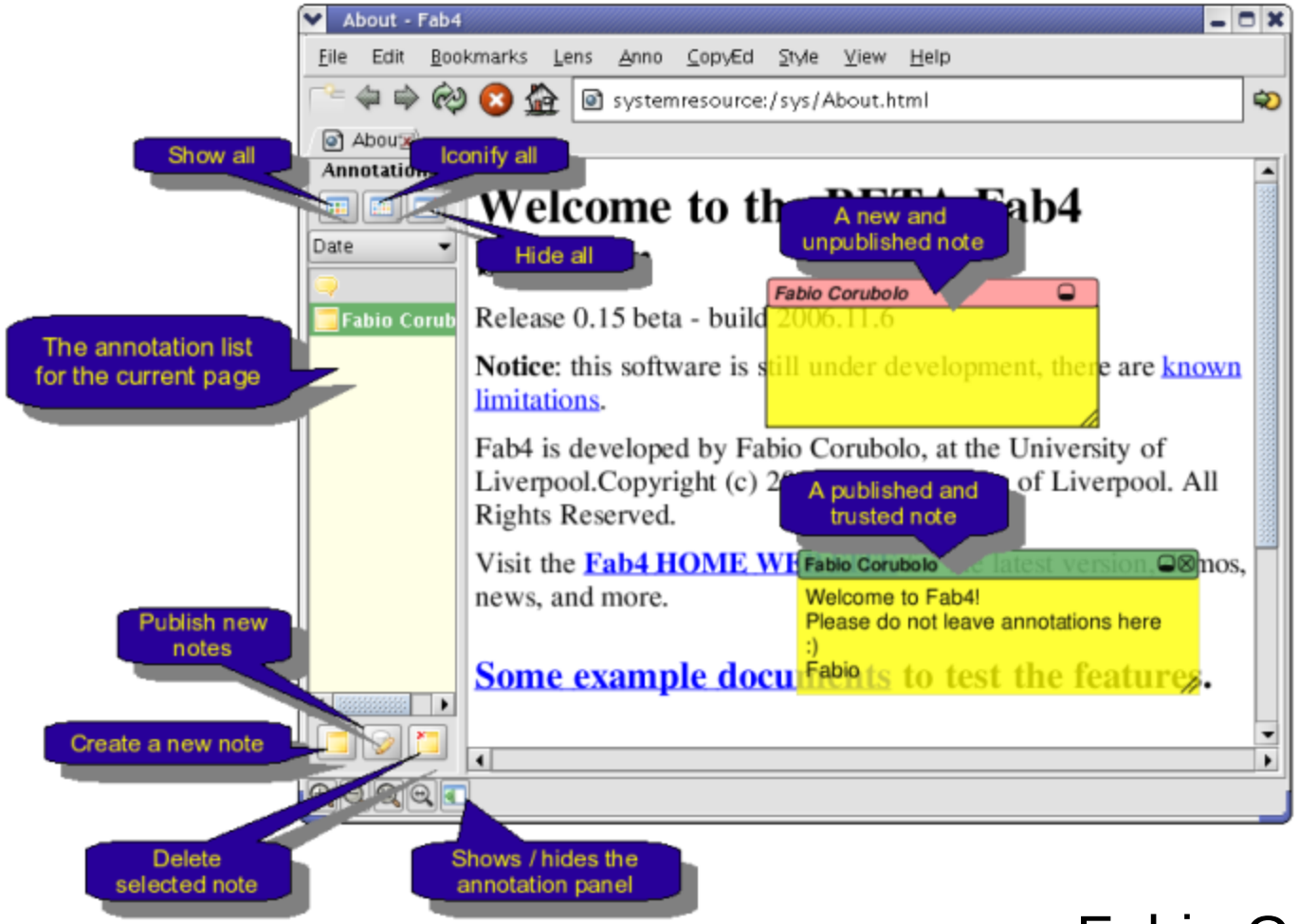
Multivalent presentation tool

- Multivalent allows data in the original encoding format to be manipulated.
- For a given data type, an adaptor (media engine) is built for the Multivalent browser
  - For example, PDF or Word
- Multivalent services can automate required processes:
  - Format identification, validation, transformation (e.g. correct invalid files)

Combined Emulation/Migration approach

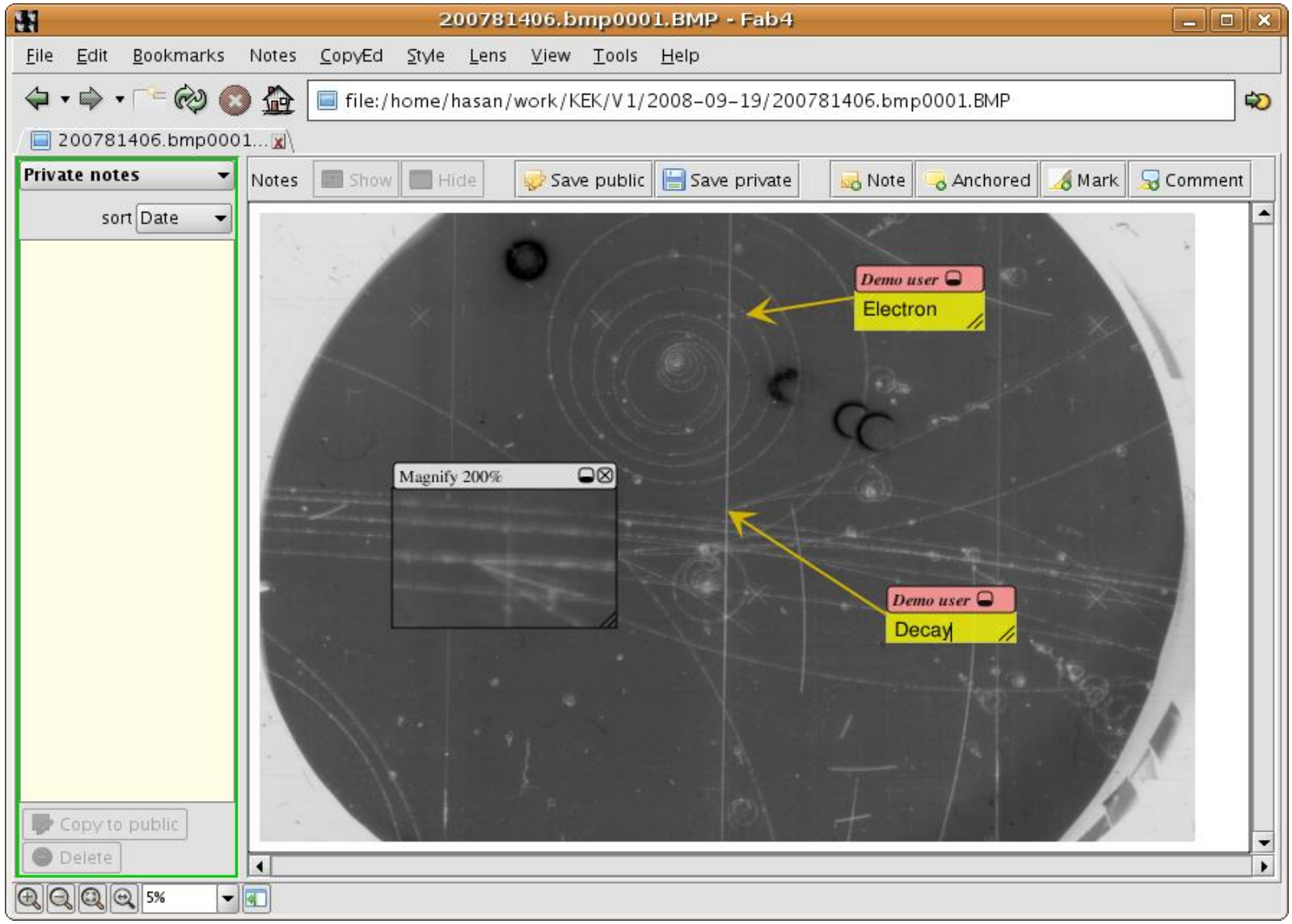
- The Multivalent technology (Java) and the media engine are archived as an iRODS collection
  - Emulation consists of supporting the original operations for manipulating the digital entity
    - We can view documents from the original bitstream
    - We can introduce new functionalities to legacy documents (e.g. magnifying lens to MacWrite 1983 documents)
  - Migration consists of porting Java virtual machine to a new system as needed
  - The *digital entity remains unchanged*, while making it possible to apply new operations

FAB4



Fabio Corubolo

Fab4


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%

Data Discovery

- Important to ensure all contextual information preserved and discoverable.
- Semantic information needs to be maintained (domain expert).
- Important to ensure external references are 'managed' (either captured or an agreement exists for their long term access).

Data Discovery (Cheshire)

- Discovery and retention are related!
- Cheshire digital library system can be used to integrate discovery and analysis in the iRODS or SRB environments
  - Cheshire processing workflows can be used to combine processes of IR, association rule mining, Semantic Web, text mining
- Understanding and generating digital ontologies can be used to aid discovery
  - Digital ontologies along with user-defined ontologies can be used in the semantic grid context

Current State

- Production version of Multivalent now released and widely used.
- Common media formats supported already.
- Production version of Cheshire digital library now released and in service.
- Work integrating Cheshire/Multivalent into iRODS now taking place.
- Further research needs to be done for application to scientific/engineering domains.
- Use for science/engineering domains is a realistic near-term goal!

References

- Multivalent:
  - http://multivalent.sourceforge.net
- Cheshire:
  - http://cheshire3.sourceforge.net
- IRODS:
  - https://www.irods.org