



**Promoting Open Standards for Digital Repository
Infrastructures:
case study examples and challenges**

Flavia Donno
CERN

P. Fuhrmann, DESY , E. Ronchieri, INFN-CNAF

OGF-Europe Community Outreach Seminar
Digital Repositories - Interoperability Using Grid Technologies

OGF23 - 5 June 2008, Barcelona, Spain





Promoting Open Standards for Digital Repository
Infrastructures:
archival case study examples and challenges

Flavia Donno
CERN

P. Fuhrmann, DESY , E. Ronchieri, INFN-CNAF

OGF-Europe Community Outreach Seminar
Digital Repositories - Interoperability Using Grid Technologies

OGF23 - 5 June 2008, Barcelona, Spain



Digital Repositories

The term **digital repository** is used with different meanings depending on the scope, the environment, and the community using it.

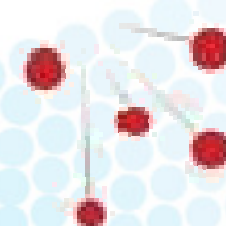
Digital Repository Infrastructure: a managed (distributed) storage system with content deposited on a personal, departmental, institutional, national, regional, or consortia basis, providing services to designated communities (Joint Information Systems Committee (JISC) Digital Repositories Program – Digital Repository Review, February 2005)



eprints

DSpace™

fedora



Digital Repository Infrastructure: Importance of standards

Modelling the international network of repositories independent of specific operation domains, and developing frameworks for interaction is very important.

It enables crossover of technologies, sharing experience across domains, collaborative development work.

External service providers can exploit the combined network of institutional repositories on a global scale:

- Cloud storage services (Nirvanix, Amazon S3)
- Indexing of metadata and content (Google Book Search)
- Data Management and Preservation services (RepliWeb/RMFT, file transformation and normalization, etc.)

Best practices, open architectures and (de-facto) standards enable integration, interoperability and collaborative work.

Digital Repository Infrastructure: Responsibilities

- **Ingestion, access and presentation**

- Data Acquisition
- Data Sharing
- Transformation services

- **Security**

- Authentication and Authorization
- Authenticity
- Privacy
- Vulnerability



- **Data Management**

- Metadata extraction, collection, manipulation
- Content and metadata management
- Workflow management

- **Storage Management**

- Space Management
- Copies Lifecycle
- Media Management
- Data Mirroring and Replication

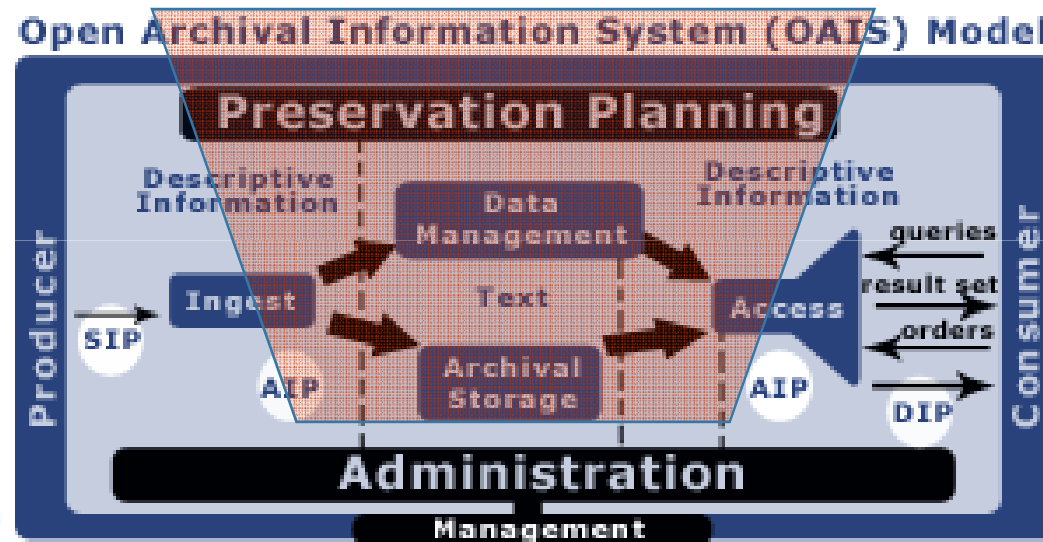
- **Preservation and Curation**

- Extend the usable life of machine-readable computer files
- Protect data from media failure, physical loss, and obsolescence.

Open Archival Information System Model (ISO)

It is a popular reference Model. It allows for the definition of layered architectures identifying the main services

Strong standardization efforts for Ingest, Access, Preservation, Administration Sustainability



Data and Storage Management standardization and best practices in distributed environments and Grid can be improved ...

Relevant Projects

- **CASPAR** – (EU-FP6) It is dedicated to preservation issues in repositories. Its goal is to search, implement, and disseminate innovative solutions for digital preservation based on the **OAIS reference model**
- **Digital Preservation Europe** - Addresses the need to improve coordination, cooperation and consistency in current activities to secure effective preservation of digital materials, focusing on authenticity and auditing issues.
- **SHERPA DP** - It is dedicated to preservation issues in repositories, developing strategy and business models for moving institutional repositories from project funding to sustainability.
- **JISC Digital Repositories Program** - It support research on institutional repositories and digital preservation.
- **Digital Curation Center** - It is in the UK, and is developing auditing and certification processes for trusted repositories.
- **RLG, NARA** - Digital Repository Certification Task Force that has established general certification requirements applying to any type of digital repository.

Available standards for description and preservation

OAIS - Open Archive Information Standard

AIP - archival information package

SIP - submission information package

DIP - dissemination information package

ISO standard - Producer-Archive Submission

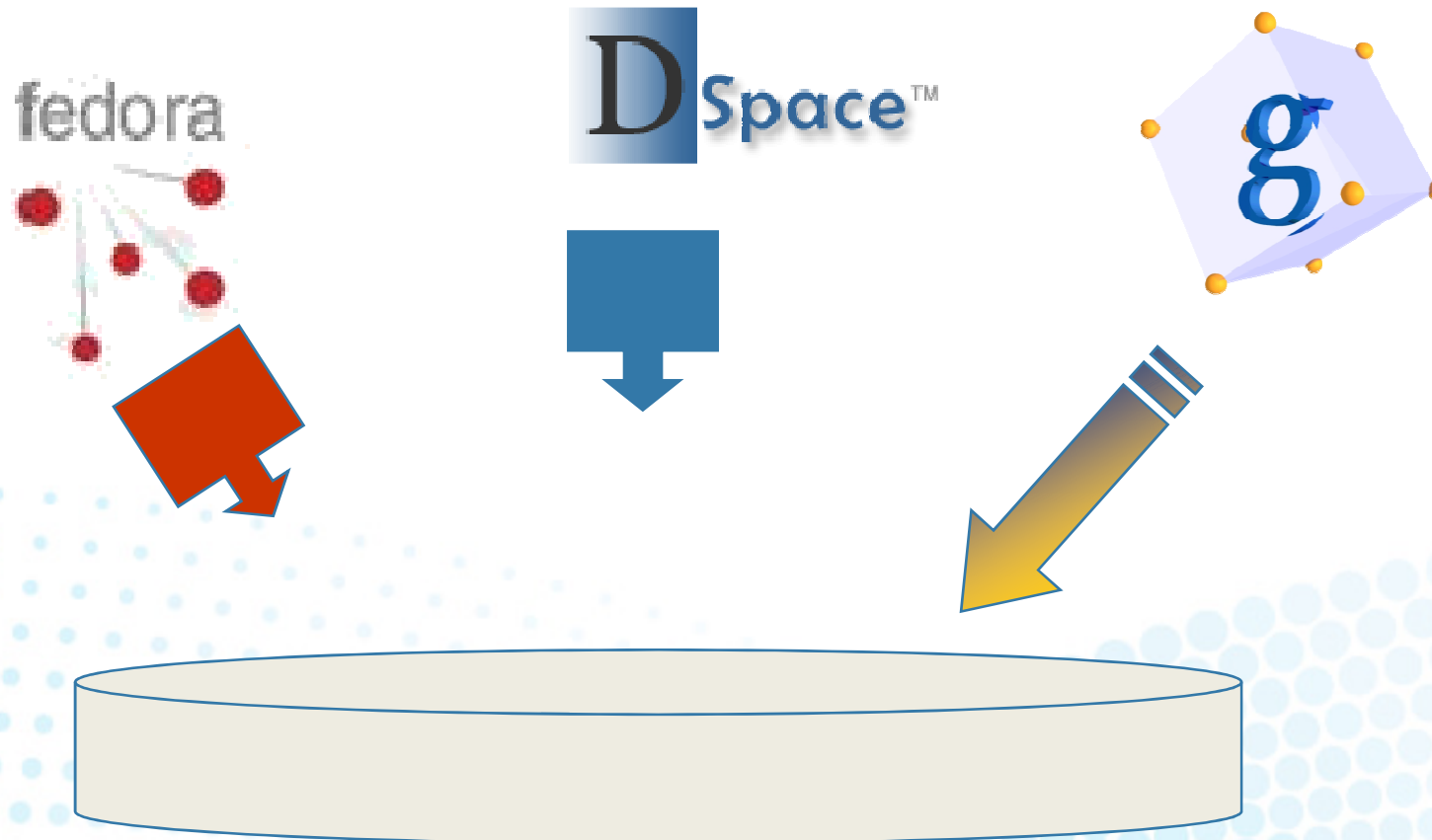
METS - Metadata Encoding Transmission Standard

PREMIS – Data Dictionary for Preservation

Dublin Core Metadata Element Set - Its goal is the interoperability and broad applicability.

MODS - MODS metadata can be mapped into Dublin Core metadata.

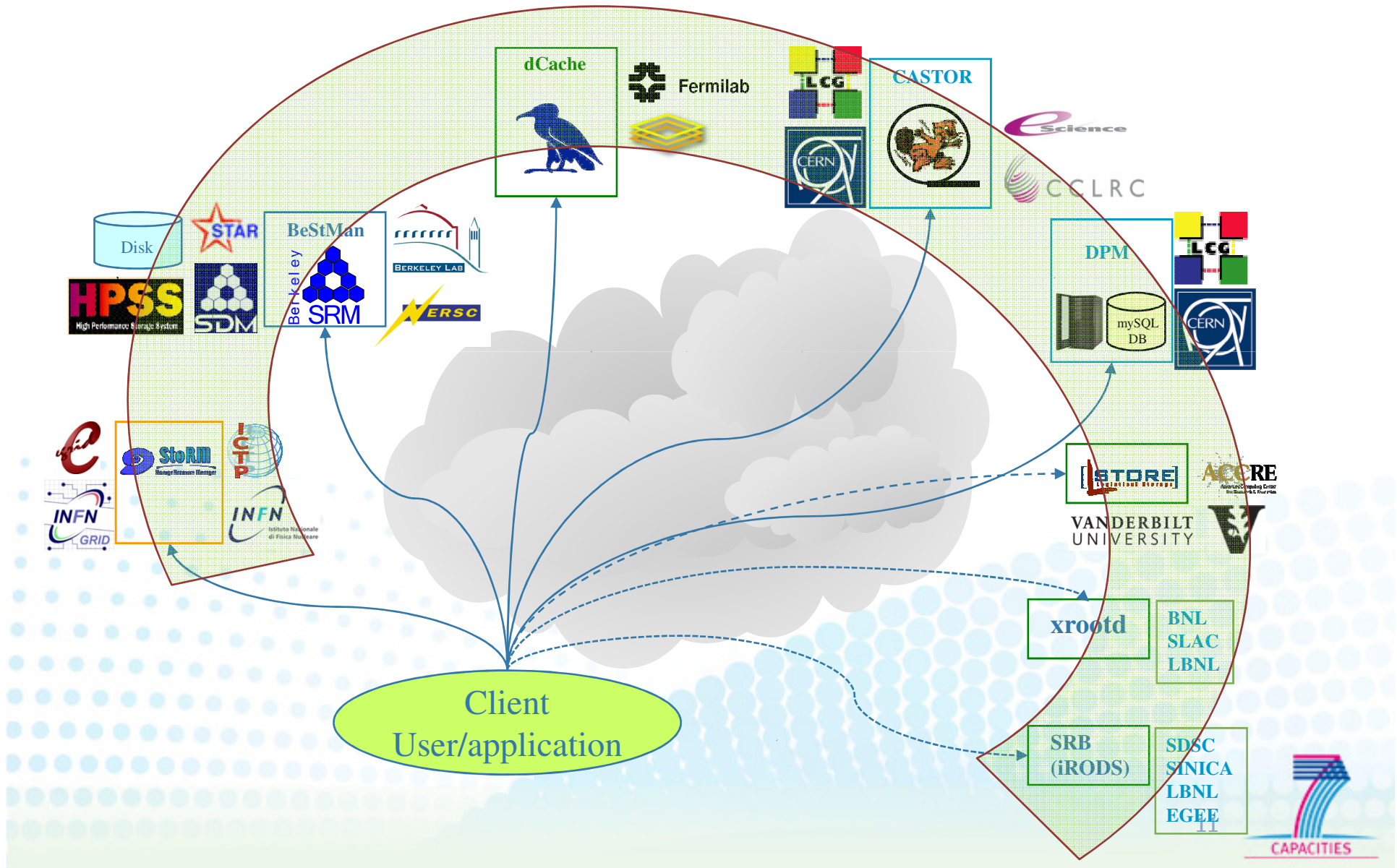
Available Frameworks with preservation/curation policies



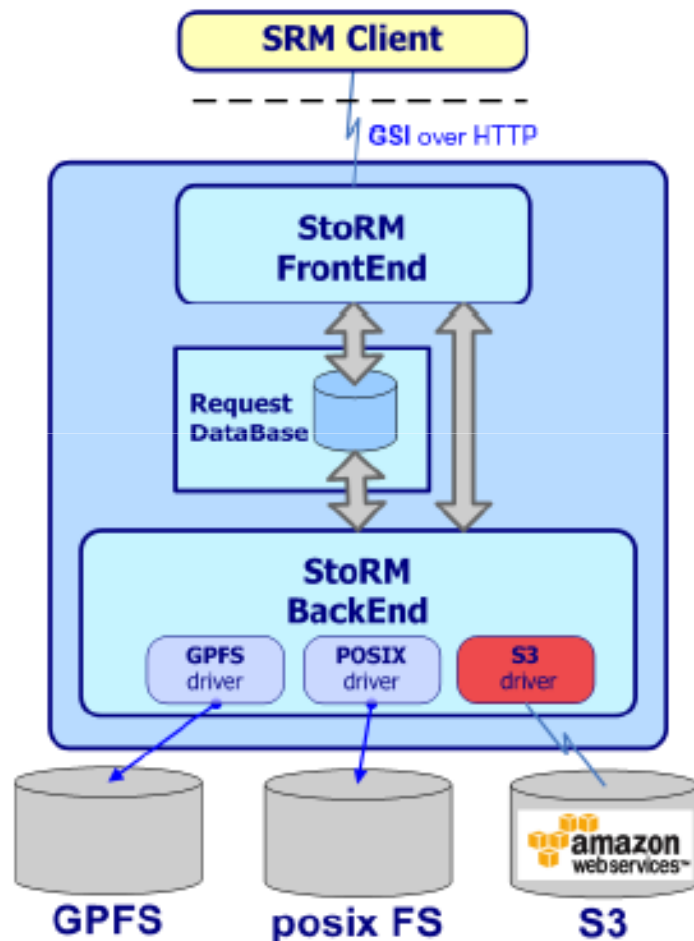
Standards for Storage: SRM

- The *Storage Resource Manager* (SRM/OGF-GSM) is an interface definition and a **middleware component** whose function is to provide **dynamic space allocation** and **file management** on shared storage components on the Grid.
 - It offers storage system independent syntax and semantic
 - It supports storage quality negotiation
 - It allows for lifecycle management of storage resources and files
 - It defines unique storage file identifiers in a Grid
 - Authorization, authentication mechanisms
 - It provides support for privacy and data authenticity
 - It allows for the negotiation of file access/transfer protocols
 - It supports optimized replication between remote storage systems

Available implementations



SRM and the clouds



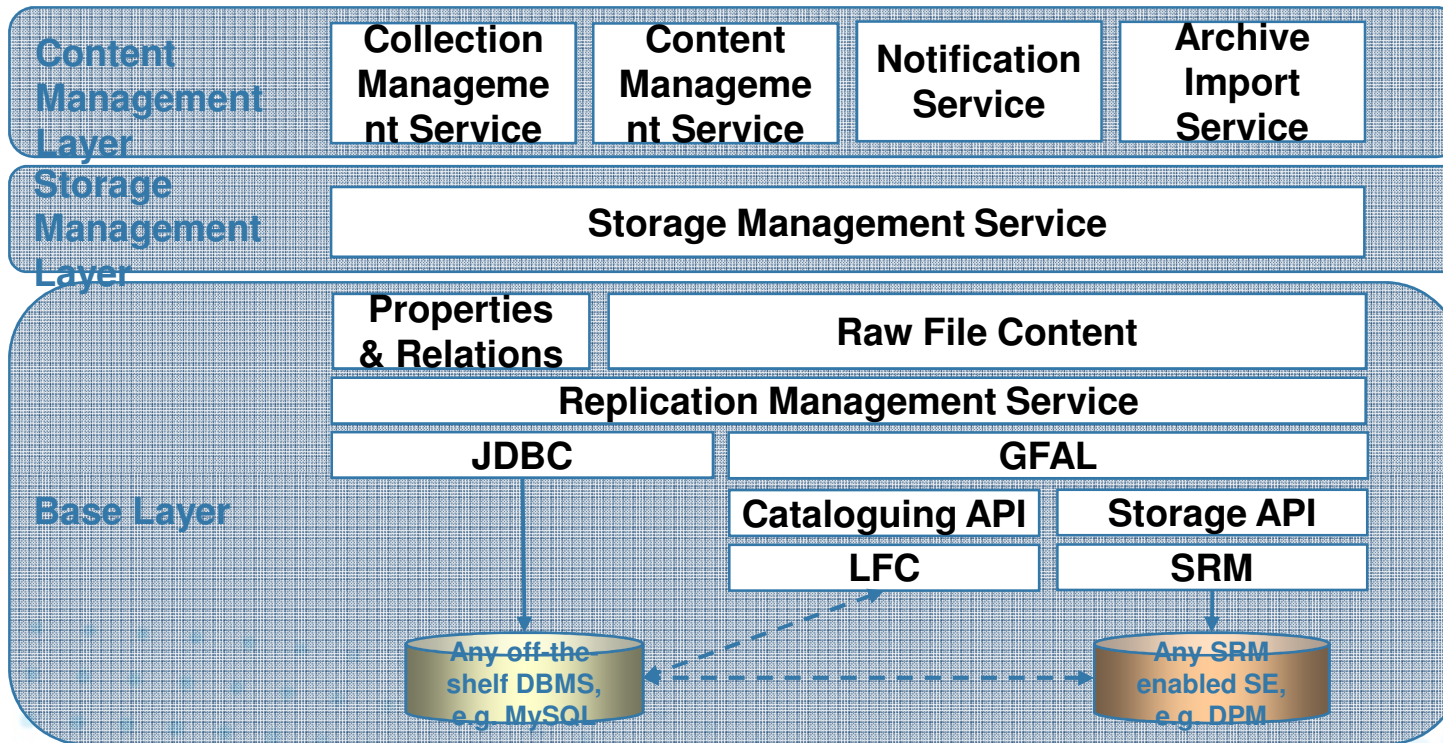
SRM interfaces to business storage services are under development

StoRM is an example

It allows for easy and transparent integration of existing SRM based environment with business providers and for a dynamic expansion of existing infrastructures

<http://storm.forge.cnaf.infn.it/>

SRM-based DR



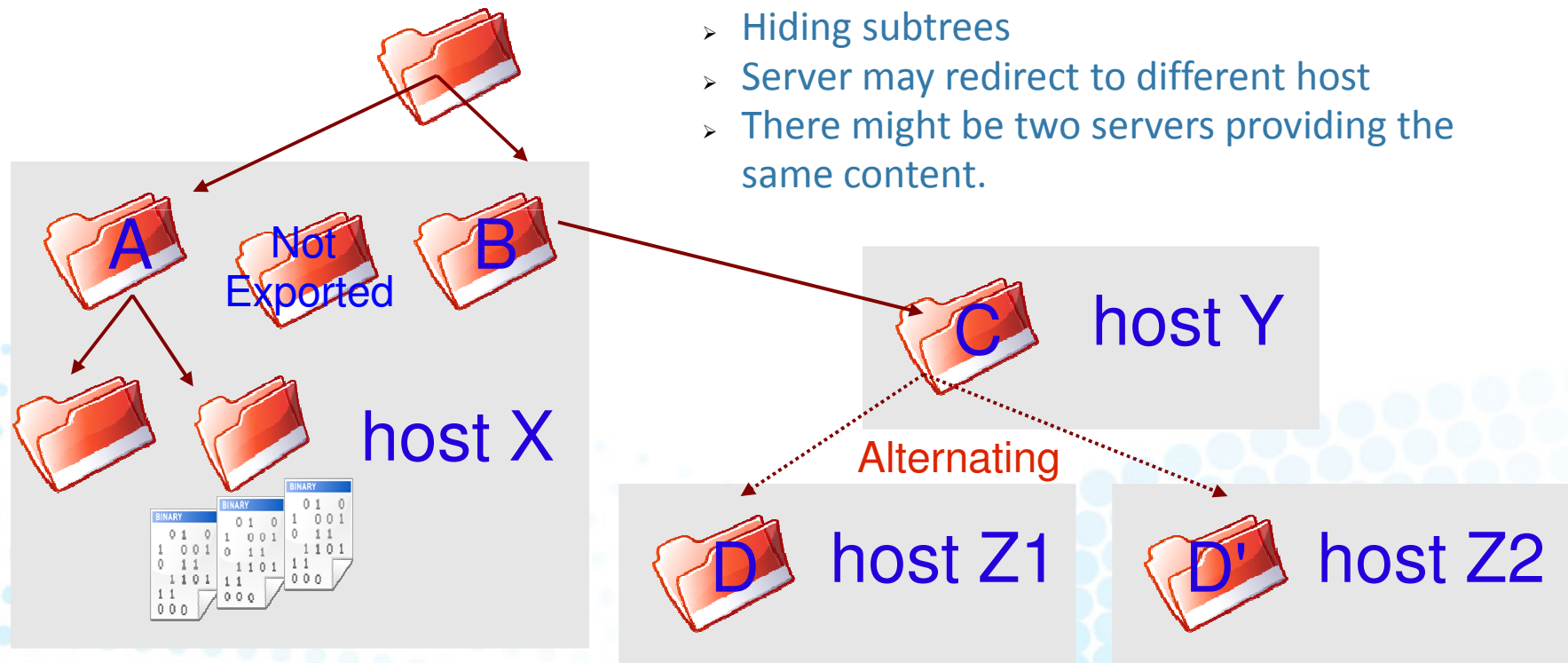
- Preservation policies enforced through replication
- Data curation will be addressed

Standards for Storage: NFSv4

- *The NFS 4 protocol is a standard since 2000. Implemented by major file (IBM/GPFS, SUN/Lustre, ...) and operating systems including Linux and Windows*
- *Specialized storage systems also offer an NFS v4 implementation: dCache now, DPM in the future*
- *Requirements to an improved (inter)net protocol*
 - Improved access and good performance on the Internet
 - Strong security, with security negotiation/delegation built into the protocol
 - Support for metadata content
 - Stateful sessions (lockd part of the protocol)
 - Enhanced cross-platform interoperability
 - Extensibility of the protocol

Attractive NFSv4 features

Arbitrary subtrees of different hosts can be exported to clients which may build their own view. So parts of the file system structure may be moved to other hosts or be replicated for performance reasons.

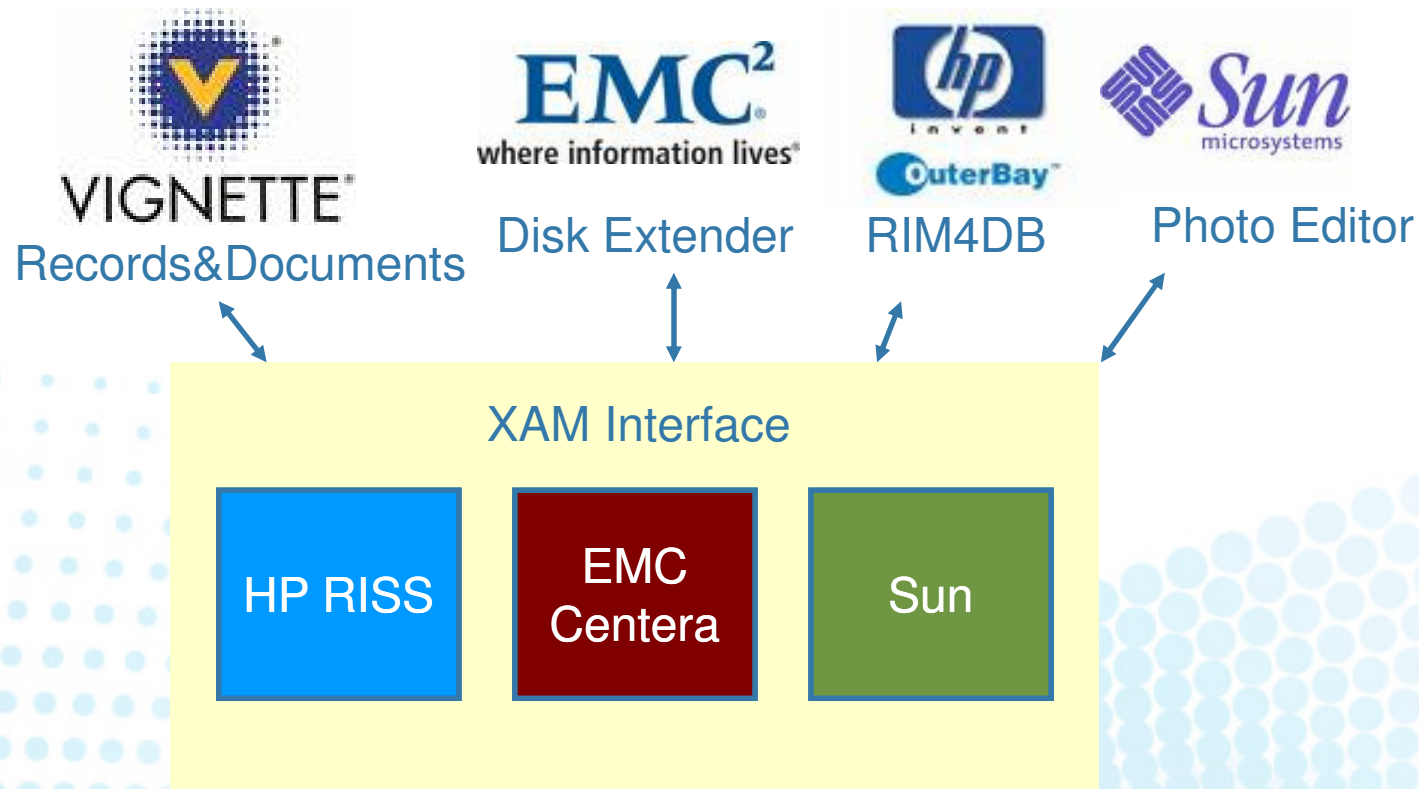


Standards for Storage: SNIA/XAM

- Metadata and content management are normally separated
- XAM allows for the storage and management of content and related metadata information in storage devices and services
- XAM specifies how metadata is represented, but not the actual metadata field names and values.
- Further work is needed to standardize metadata names and allowed values for application domains like Email, Health, and Document Management.

SNIA/XAM at work

Commercially Available Applications



Challenges

- Coordination of disciplinary specific initiatives to encourage the adoption of common practice for data archival, preservation, and curation
- Demonstration of federated specialized storage and computing infrastructures to support digital repositories. Requirements and needed interfaces
- Integration and interoperability of existing repository infrastructures, data exchange and accessibility
- Common operational procedures for networked digital repositories.
- Dissemination activities and reference fora for producers and administrators