Diversity and Interoperability of Repositories in a Grid Curation Environment

Jens Ludwig
ludwig@sub.uni-goettingen.de
SUB Göttingen

Andreas Aschenbrenner, SUB
Harry Enke, AIP
Thomas Fischer, SUB

SPONSORED BY THE

Federal Ministry
of Education
and Research

Background of WissGrid

- WissGrid is part of the German National Grid Initiative (D-Grid)
- D-Grid covers a wide range of academic disciplines and industrial partners
- WissGrid‘s objective: represent academic user interests (negotiate with alliance of computer centers and industry)

WissGrid‘s Detailed Objective

Establish long-term sustainable

- organisational and
- technical

D-Grid infrastructure for the academic world

Three areas of work (aka work packages):

- Operational model for academic grid users
- Blueprints for new community grids
- Long-term preservation of research data

The WissGrid Partners

The partners of WissGrid are representatives of five established academic grid communities:

- HEP-Grid: high energy physics
- TextGrid: humanities
- C3-Grid: climate sciences
- Medi-Grid: medicine
- AstroGrid-D: astronomy

But new communities like social sciences, bio statistics, photon sciences, etc play also a significant role in WissGrid.

Situation

Coming from a library world: How to improve the data management situtation of research data? In a variety of disciplines?

Prerequisites of communities vary considerably:

- existing large-scale data repositories vs no repository
- homogeneous data vs heterogeneous data
- immutable data vs changable and erasable data
- open access data vs personal and sensitive data
- specialized expertise since decades vs no knowledge

General Approach

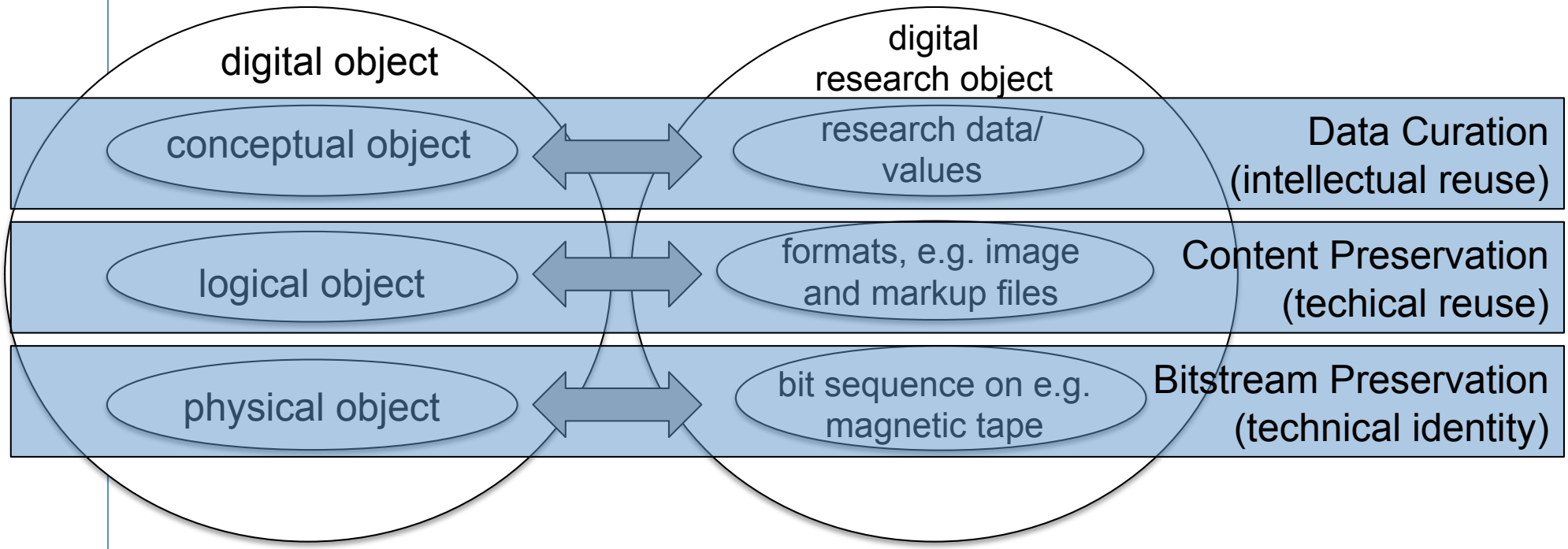
- Every community should benefit without having to adopt everything.
- Provide cross-disciplinary and generic data curation tools and offer basic data curation blueprints.
- Adapt and combine a basic set of tools for the D-Grid environment (Fedora, iRODS, DCache, JHove2, ...).
- Respect diversity of systems and foster interoperability!

Finding a Common Terminology

But before we could start:

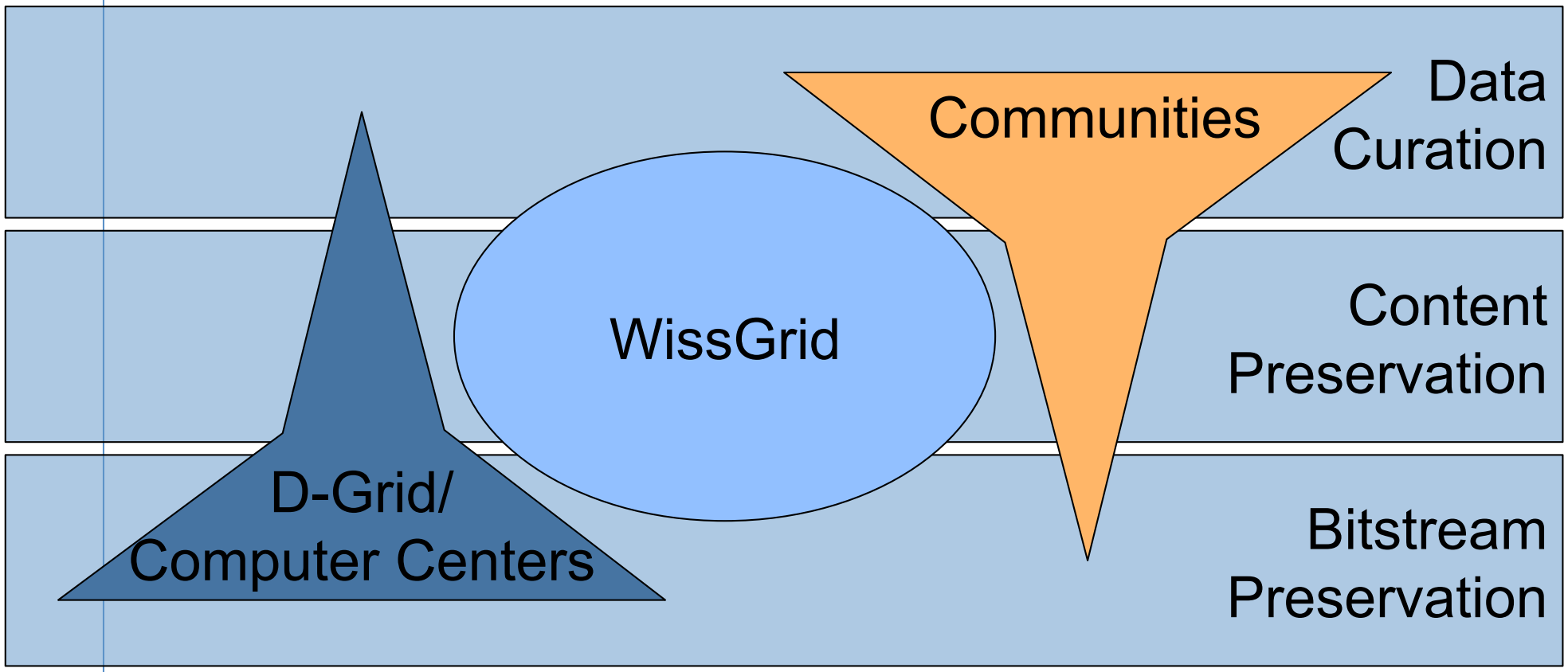
We had to settle
with a variety of disciplines
on a common terminology
and common concepts.

Three Aspects of Long-Term Preservation



Inspired by Thibodeau: Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years, CLIR 2002.

Responsibilites for Developments



WissGrid‘s Developments

For the individual levels:

- Bitstream Preservation: advocate and define requirements
- Content Preservation: adapt generic tools to D-Grid environment (JHove2 for format characterization, conversion services, ...)
- Data Curation: needs to be dealt with on a user-specific level, provide guidelines and consultancy

Encompassing all levels: Repositories

- Storage
- Technical services
- Metadata/intellectual modeling

Differences in Interest of the Communities

We surveyed our target communities. All development plans were welcomed. The „worst“ result was that of 11

- only 3 wanted to adopt fully and
- only 4 wanted to adopt partially

a development.

Not surprising: Established communities showed

- slightly less interest in tools and
- clearly less interest in repositories

than new communities.

(Only HEP says that they have everything they want.)

Grid-Repository Integration Patterns

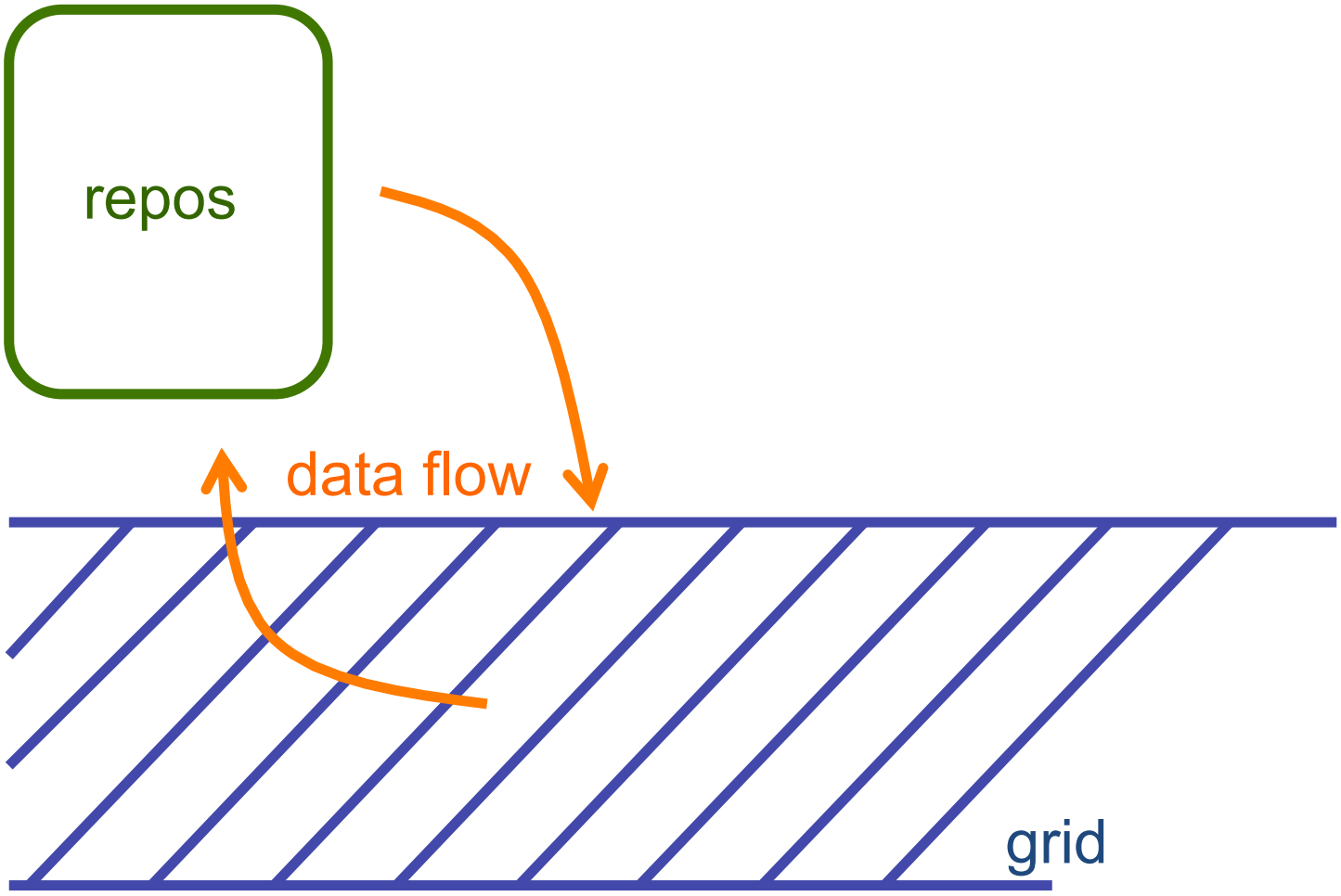
Core of WissGrid's agenda:

- Integration of repository systems into grid research environments of the communities
- by providing standard software packages for different repository/curation purposes.

We see five variants:

1. Repositories as archive backends for the grid
   (resp. a compute grid for repositories)
2. Data grid as repository storage
3. Virtualization of repositories (aka federation)
4. Embedding of repositories in scientific workflows
   (trivial, omitted)
5. Repository modules integrated in grid technologies
   (like on-the-fly virtual repositories, too complex, omitted)

Grid-Repository Pattern: Compute Grid (Repositories as Archive Backends for the Grid)

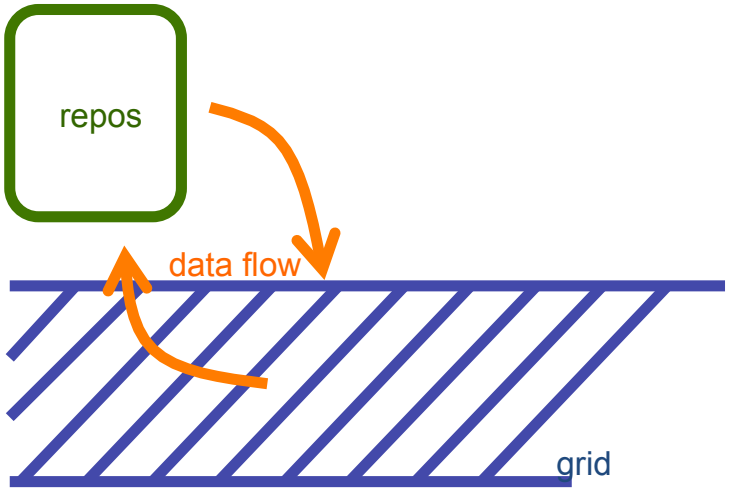


Grid-Repository Pattern: Compute Grid (Repositories as Archive Backends for the Grid)

- Scientific applications process data in the grid environment and archive it in the repository

or similiar

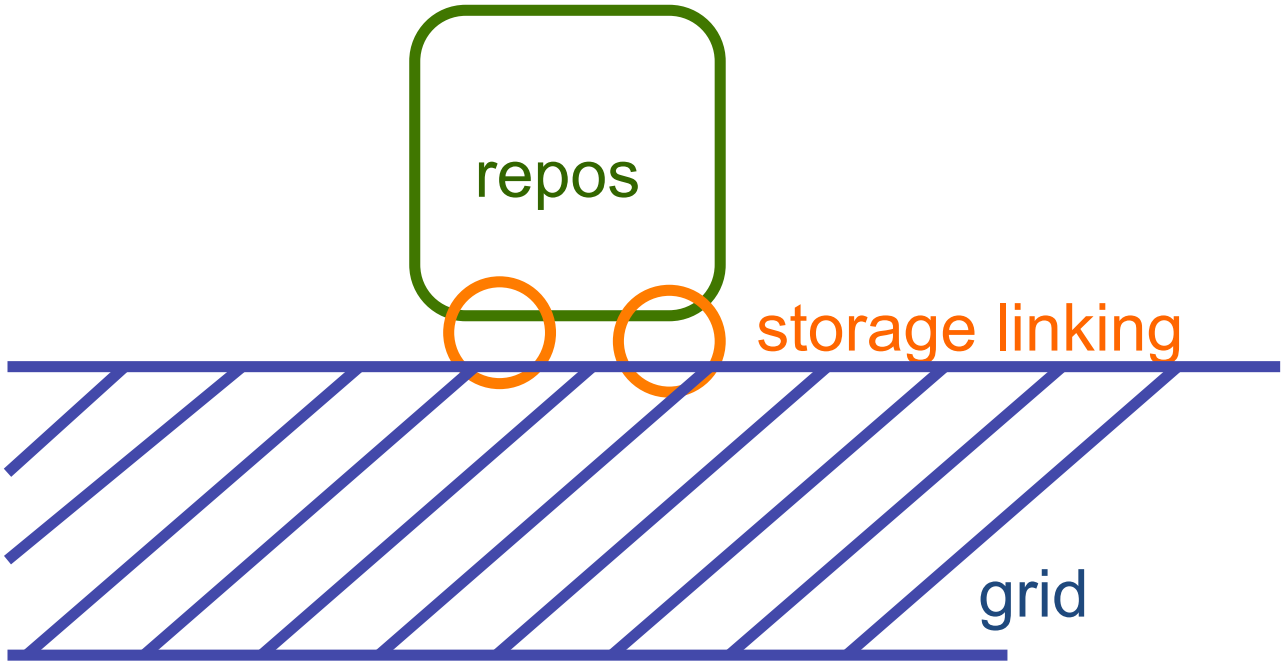
- digital objects are managed in digital repositories and the grid is used for computation.



Considerations:

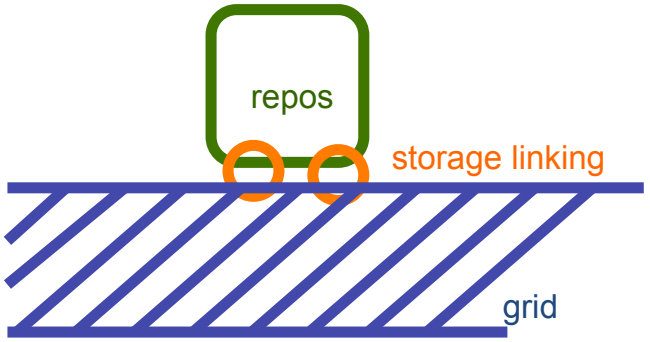
- Repository needs standard (grid) interfaces for the data to be searched, extracted, written, ...
- Mapping of rights between grid and repository
- Data to the services vs services to the data

Grid-Repository Pattern: Storage Grid



Grid-Repository Pattern: Storage Grid

- Digital objects are managed in digital repositories and the grid is used for storage
- Rationale: External storage provider is more efficient (and maybe offering replication, scalability, integrity checks?)

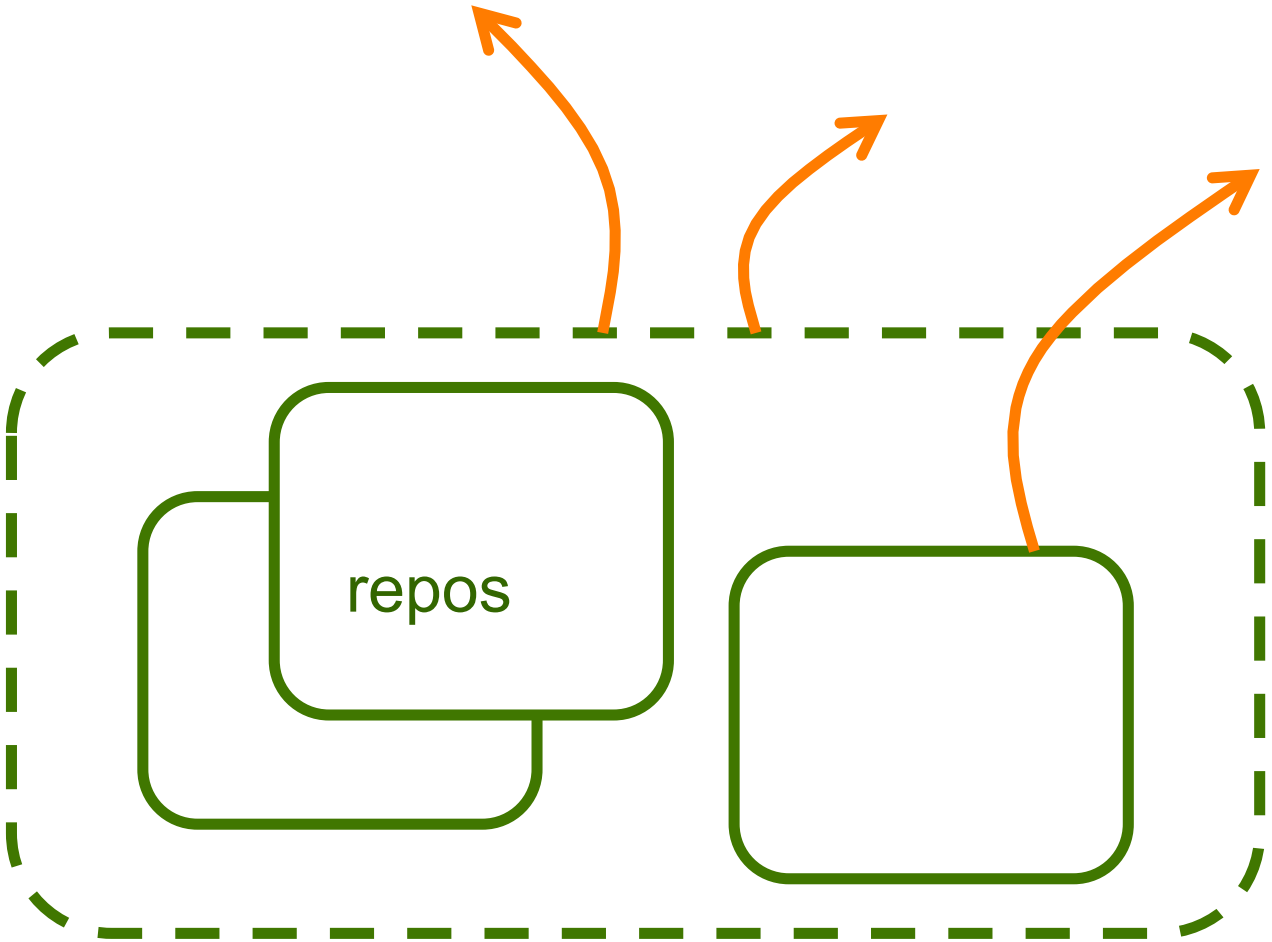


Considerations

- Security for data in the grid, bitstream preservation, SLAs (Service Level Agreements), mapping of rights between grid and repositories

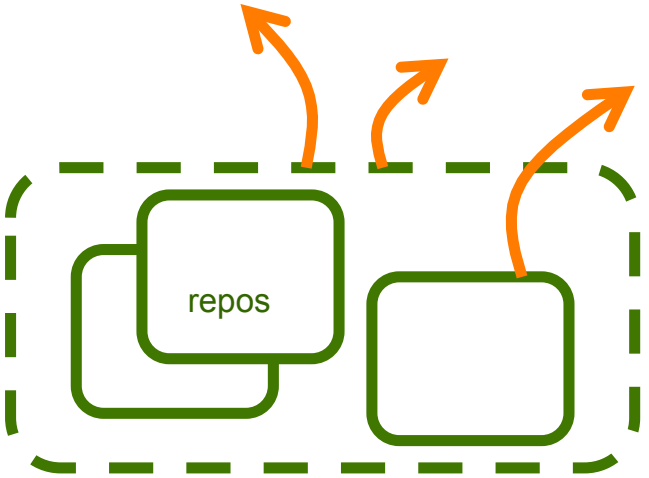
- Data could/should be accessible directly through grid mechanisms (syncronisation and security issues).

Grid-Repository Pattern: Federation



Grid-Repository Pattern: Federation

- Federation of distinct data sources that (already) exist within a single community or multiple communities.
- Not grid in a narrow sense (virtualization of services instead of hardware resources)



Considerations:

- uniform metadata profiles, common interfaces, central services, ...
- very discipline specific

Main Implementation Tasks

Repositories as archive backends:

- use iRODS as repository
- extend iRODS with CQL/OpenSearch, OAI-ORE export, OAI-PMH, grid storage interfaces (SRM), bitstream preservation
- rights management (GSI, Grid Security infrastr.)

Storage Grid:

- use Fedora/iRODS (adapting community efforts)
- rights management (GSI, x.509 certificates)

Federation:

- no implementation, only guidelines
- but: federation of repositories from above packages might be easier...

Thank you!