

## Digital Archive Federations – DAM-LR and CLARIN

Peter Wittenburg

Daan Broeder

Sven Stromqvist

Remco van Veenendaal

**Jacqueline Ringersma**

---

## Max Planck Gesellschaft

78 research institutes (Germany)

3 outside Germany:

2 Italy (art)

1 The Netherlands (psycholinguistics)

The study of mental processes involved in language production, language comprehension and language acquisition, as well as the relation between language, thought, and culture



## Distributed Access Management for Language Resources



Max-Planck Institute for Psycholinguistics



Institute for Dutch lexicology, Netherlands



Lund University, Sweden



School of Oriental and African Studies, UK

# Why DAM-LR? Language resource archives

Trend towards **centralized** storage of digital language resources:  
*Language resource archives*

Documentary linguistics:

MPI (NL), SOAS (UK), AILLA (Austin), Paradise© (Sydney), LACITO (Fr.)

Computational linguistics and speech technology:

Bavarian Speech Archive (Germany)

INL-TST (Dutch language and speech technology centre)

ELDA (France)

---

# Language Resource Archives

## *Aim of Language resource archives*

Long term data preservation (traditional archiving goal)

Easy deposition

Dissemination

Enrichment

On line services, web access →

opportunities to integrate fragmented corpora

---



# Language Resource Archives

## Integrate fragmented corpora

Different archives holding resources for a particular language – or any other aggregation mechanism – provide users with a seamless domain for search and access

---

# Language Resource Archives

Benefits for the user:

*Overcoming **format differences** of resources within and across archives*

*Overcome **concept naming issues** through ontology mapping*

*(not in DAM-LR)*

**Resource discovery:** one metadata set for searching & browsing

**Access mechanism:** single user identity, single sign-on

A single system of **identifiers for referencing** “archived resources”

---



# DAM-LR project (2001 – 2007)

## Goals:

Trusted servers and services

Deep metadata for research purposes

Stable and unique resource identifiers

User management and authentication (identity federation)

---

# Federation technologies

DAM-LR federation **technology pillars**:

Integrated metadata domain

Unique persistent resource identifiers (PID)

Single user identity and single sign-on

Federation-wide authorization system

**Formal arrangements** for federation members:

National bodies accept certificates from each other with a Public Key Infrastructure

Federation members have status of Registration Authority

→ Federation is based on a domain of trusted servers and services

---

# Metadata domain

Goal of the joined **metadata domain**:

Resource discovery

Metadata browsing over a unified catalogue

Single query search over all archives

Because: Interoperability through IMDI metadata

Archives can takes part by:

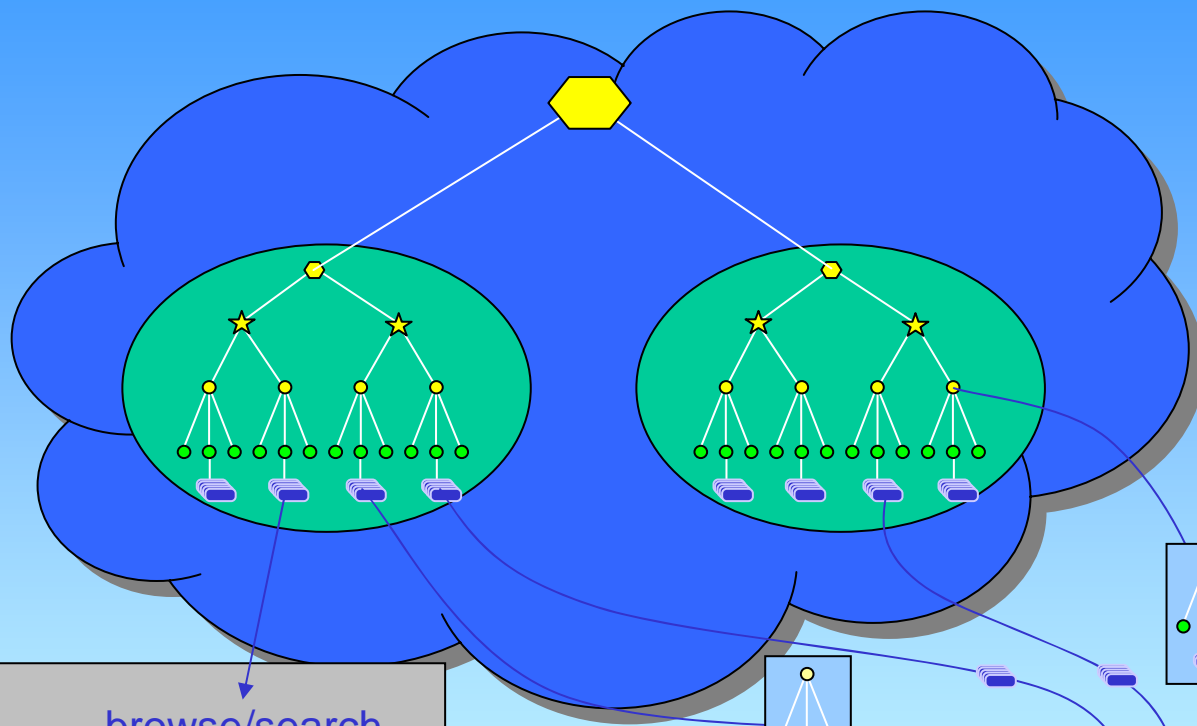
Allow harvesting by IMDI service providers:

- \* native IMDI XML data
- \* create IMDI metadata on the fly

Run IMDI archive web-applications

---

# Shared Metadata Domain



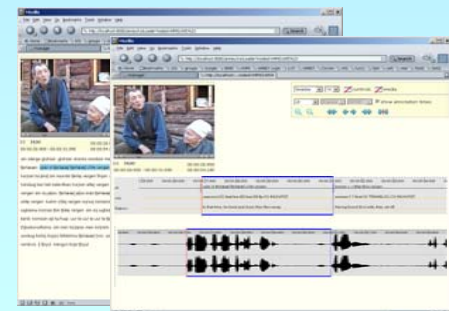
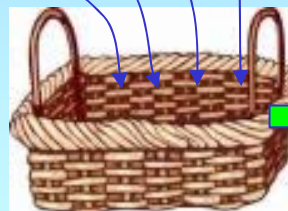
- Use IMDI metadata for resource discovery
- User selects resources from different archives
- Browsing & searching
- This is his/her private virtual temporary domain
- Possibly then again search (MD+content)
- Utilization
  - visualization/listening/...
  - comparison
  - statistics
  - ...

browse/search & select

authenticate authorize

play

web-browser & plug-ins





# Persistent Identifiers - PID

Primary goal: Avoid link rot

Separate object identity from its location

Give every resource a unique persistent identifier: PID

Every PID associated with one (or more) URL (s)

Resolving process built into applications

or available through plug-ins in web browsers.

This comes at a cost:

Added layer of infrastructure must be managed

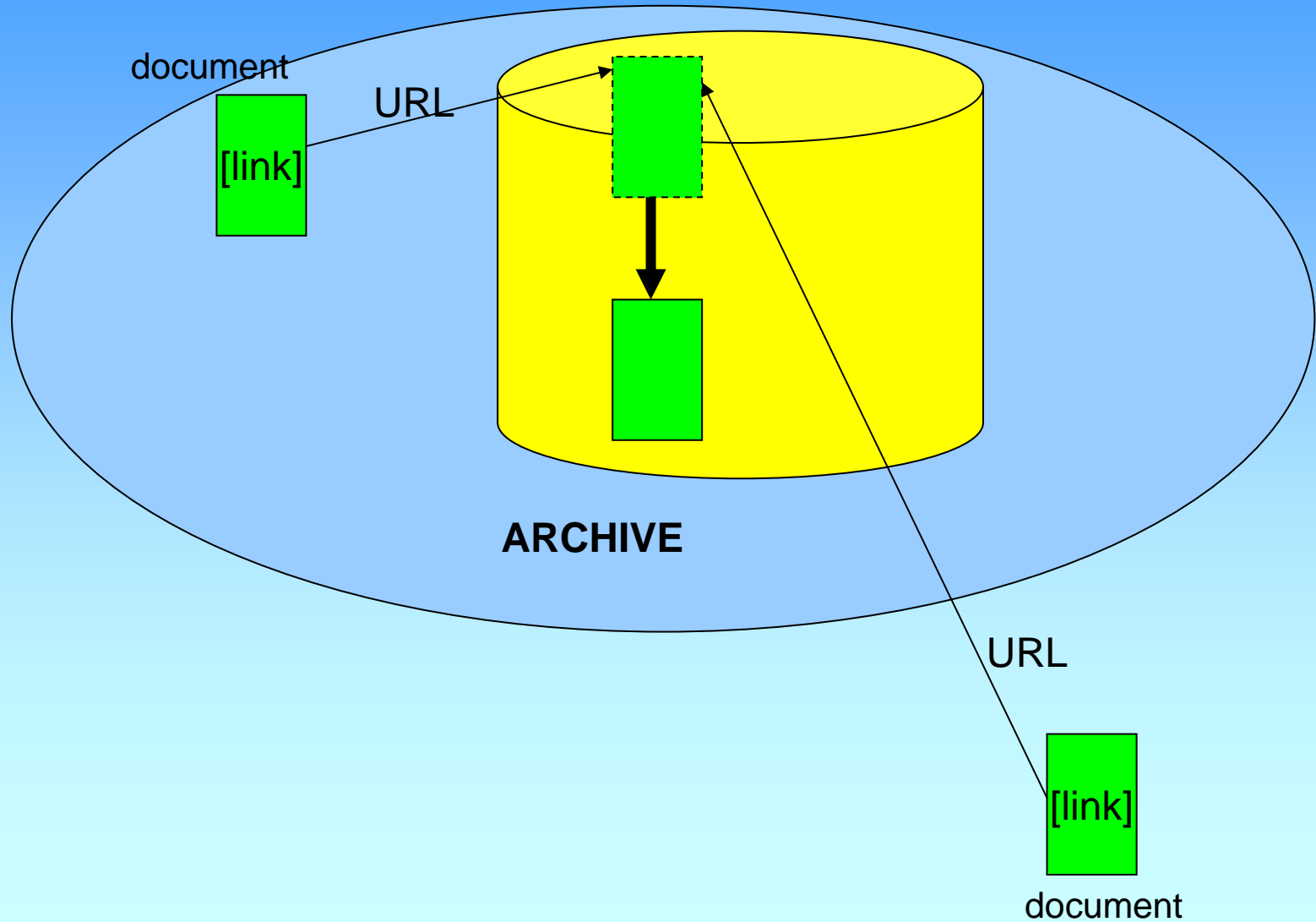
Update the PID/url info when moving the object

Resolver service must run with high availability

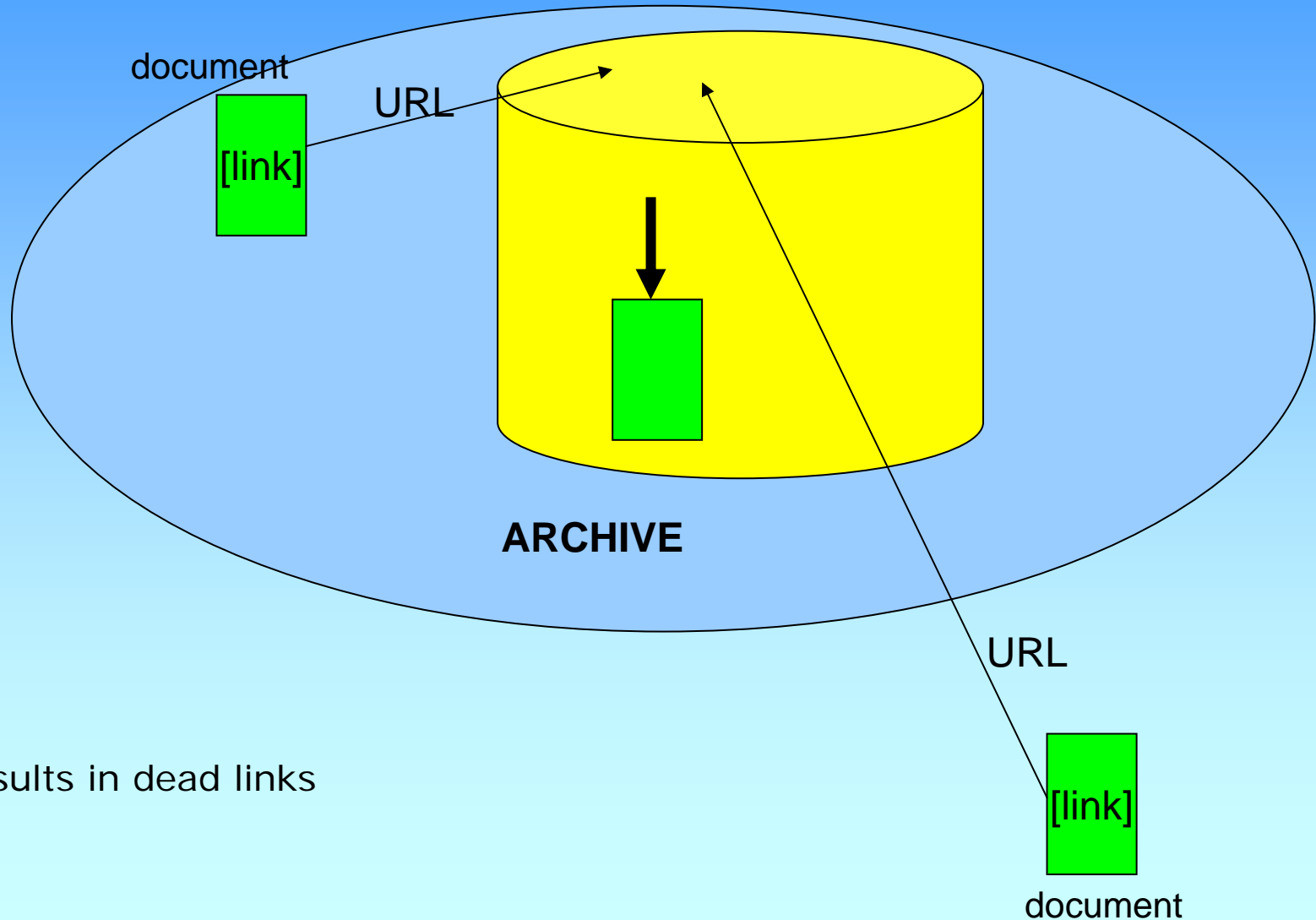
Must be very sure that this can be handled by our archives (long term).

---

# Unique Identifiers - PIDS

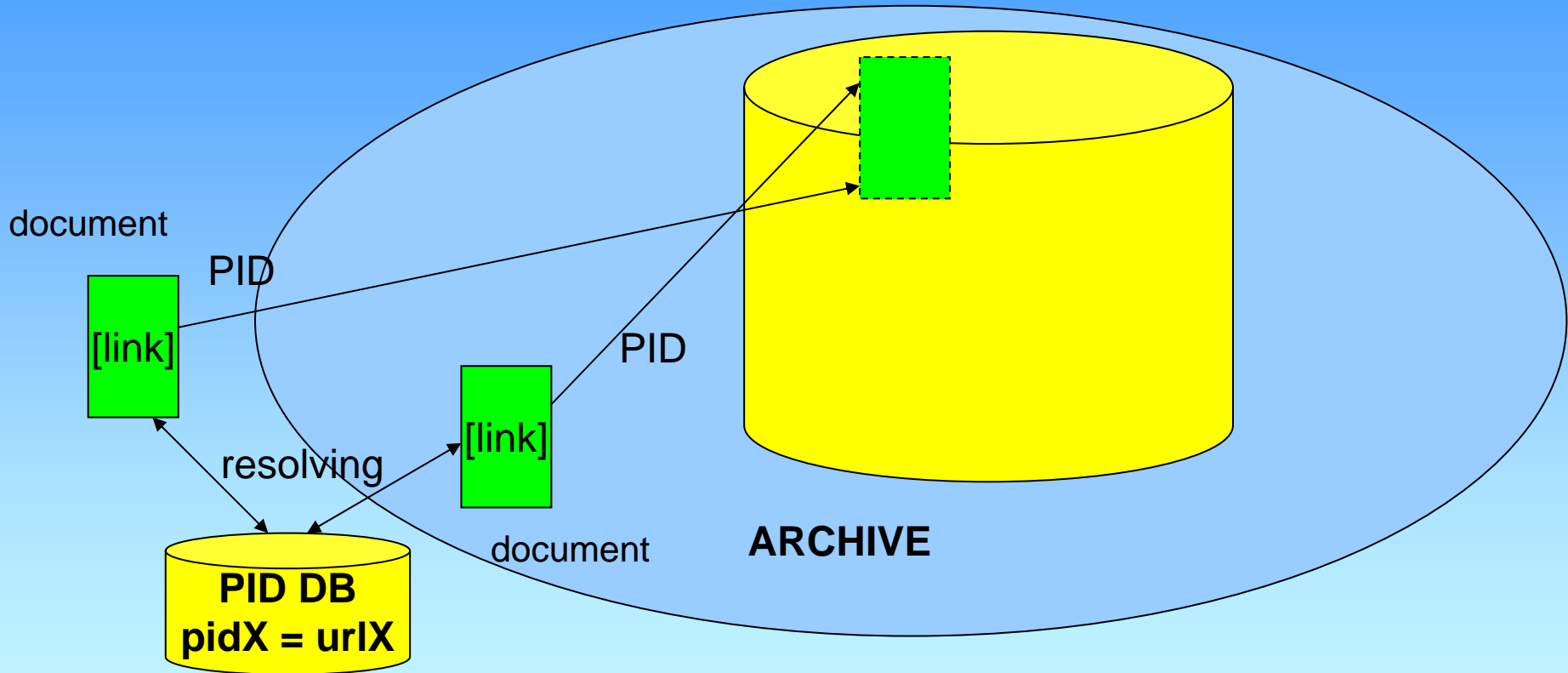


# Unique Identifiers - PIDS

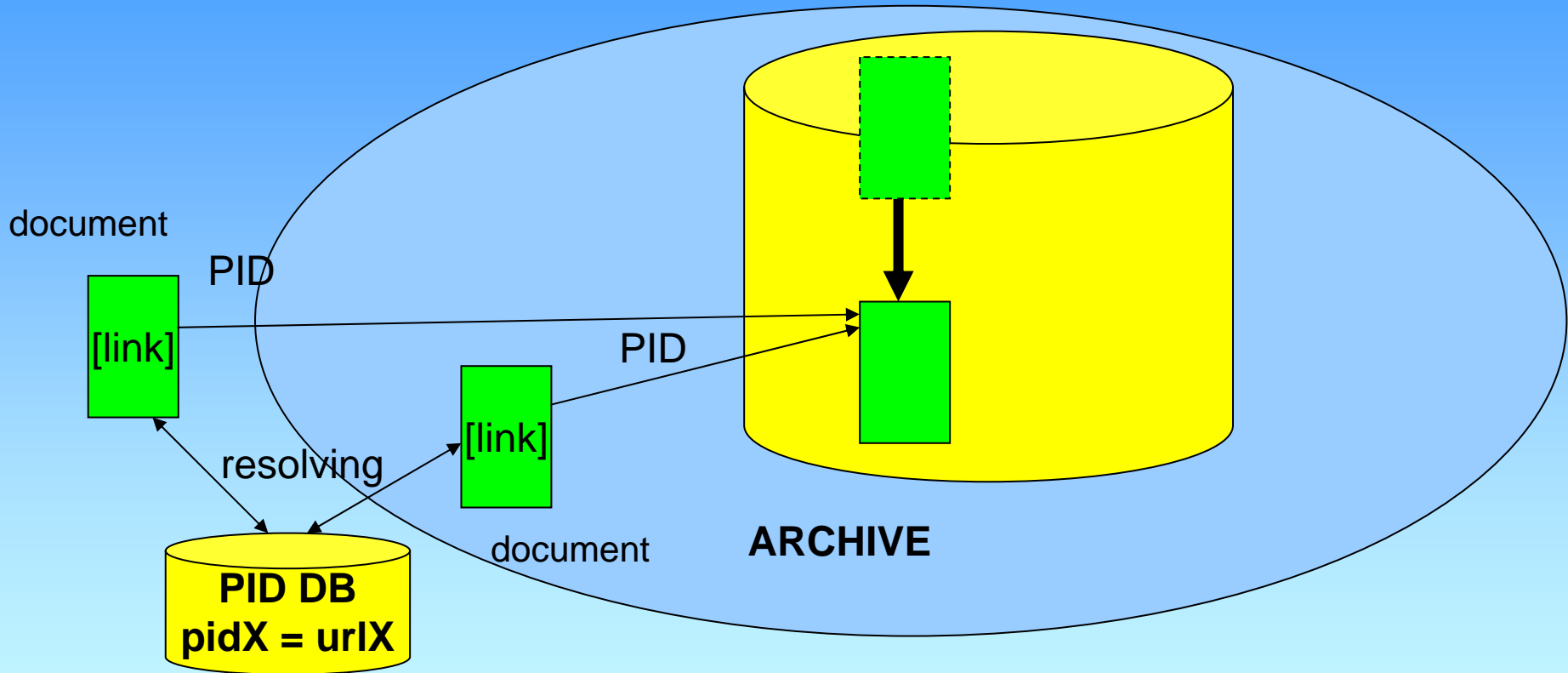


Move results in dead links

# Unique Identifiers - PIDS



# Unique Resource Identifiers - PIDS



PID is maintained when moving the object → no dead link



# DAM-LR uses Handle System from CNRI

Corporation for National Research Initiatives: [Handle System](#)

General purpose distributed information system

Open set of protocols

Provides efficient, extensible, and secure [identifiers](#) and [resolution services](#)

Resolves those handles into the information necessary

to [locate \(& access\)](#) the resources.

Can be changed as needed to reflect the current location of the identified resource without changing its identifier,

[the name of the item persists](#) over changes of location

---

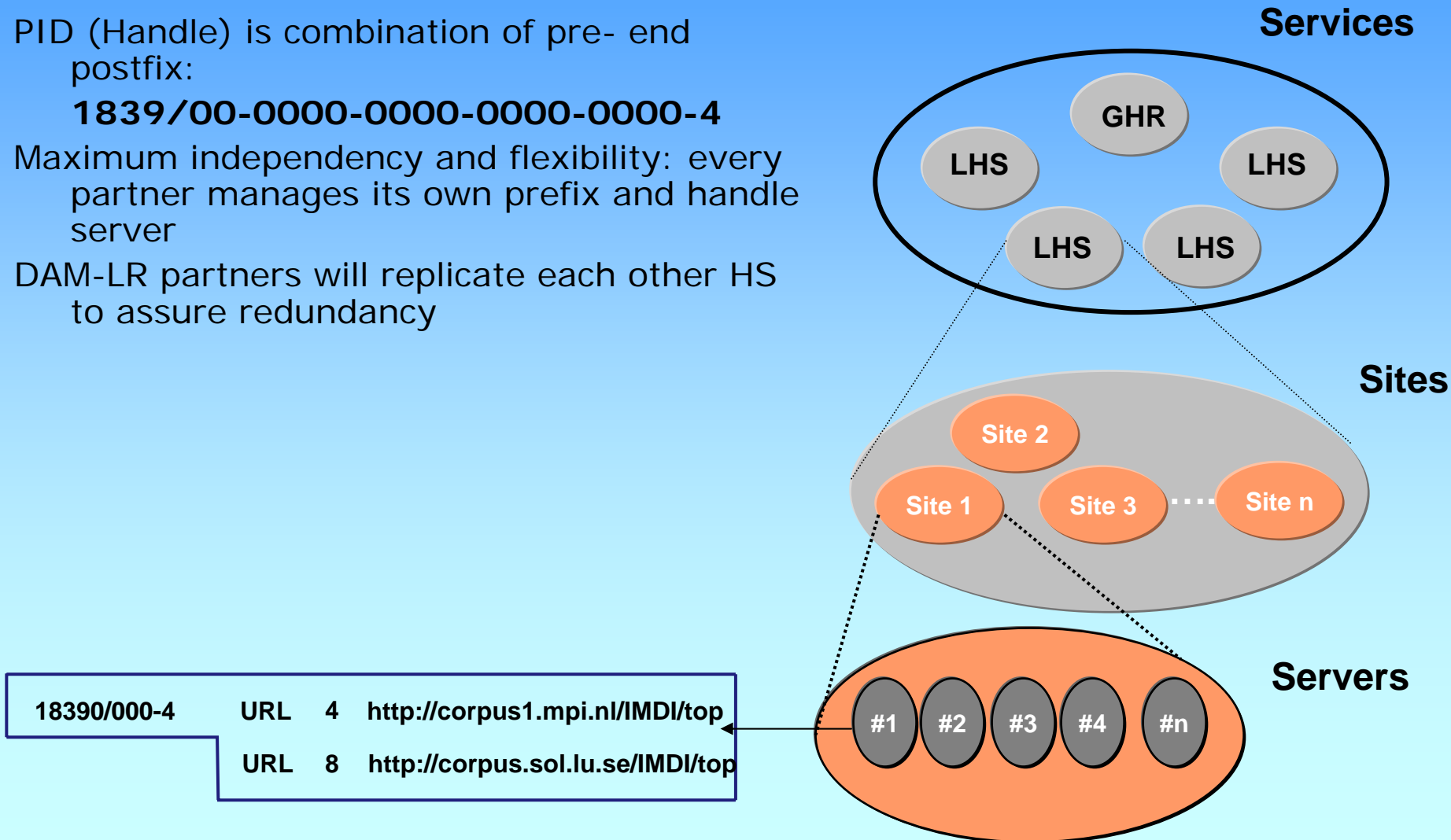
# DAM-LR Handle system from CNRI

PID (Handle) is combination of pre- end postfix:

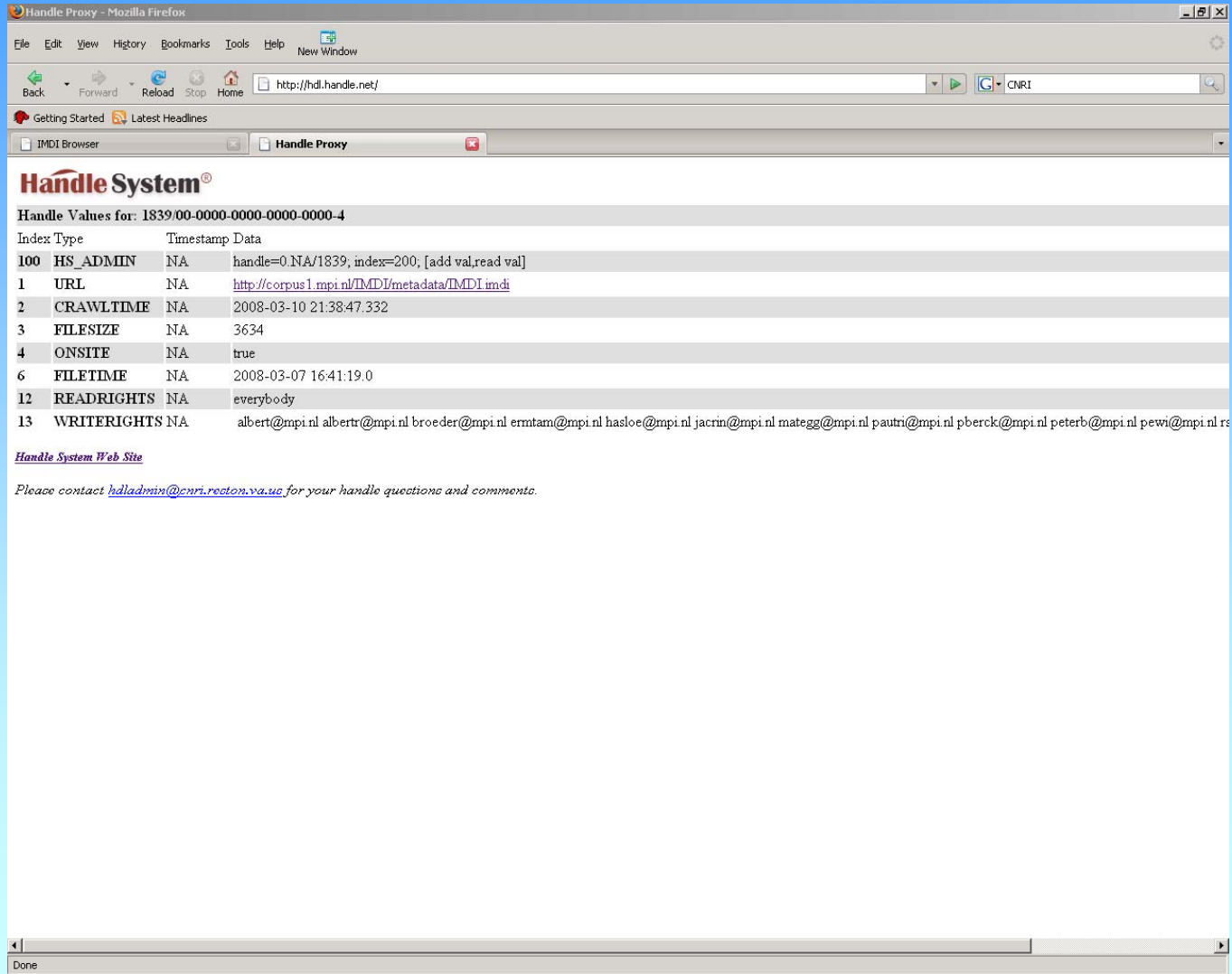
**1839/00-0000-0000-0000-0000-4**

Maximum independency and flexibility: every partner manages its own prefix and handle server

DAM-LR partners will replicate each other HS to assure redundancy



# DAM-LR Handle system from CNRI



Handle Proxy - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Back Forward Reload Stop Home http://hdl.handle.net/ CNRI

Getting Started Latest Headlines

IMDI Browser Handle Proxy

## Handle System®

Handle Values for: 1839/00-0000-0000-0000-0000-4

Index	Type	Timestamp	Data
100	HS_ADMIN	NA	handle=0.NA/1839; index=200; [add val,read val]
1	URL	NA	<a href="http://corpus1.mpi.nl/TMDI/metadata/TMDI.mdi">http://corpus1.mpi.nl/TMDI/metadata/TMDI.mdi</a>
2	CRAWLTIME	NA	2008-03-10 21:38:47.332
3	FILESIZE	NA	3634
4	ONSITE	NA	true
6	FILETIME	NA	2008-03-07 16:41:19.0
12	READRIGHTS	NA	everybody
13	WRITERIGHTS	NA	albert@mpi.nl albert@mpi.nl broeder@mpi.nl ermtam@mpi.nl hasloe@mpi.nl jacrin@mpi.nl mategg@mpi.nl pautri@mpi.nl pberck@mpi.nl peterb@mpi.nl pewi@mpi.nl rs

[Handle System Web Site](#)

Please contact [hdladmin@cnri.reston.va.us](mailto:hdladmin@cnri.reston.va.us) for your handle questions and comments.

Done

# User Authentication & Authorization

## Sharing users

- Single user identity

- log-in only once when working with resources from multiple archives

- user management and authentication should be left to home institute

## Sharing resources

- origin institute sets access rules to resources

- access rules from origin institute should be respected for copied resources

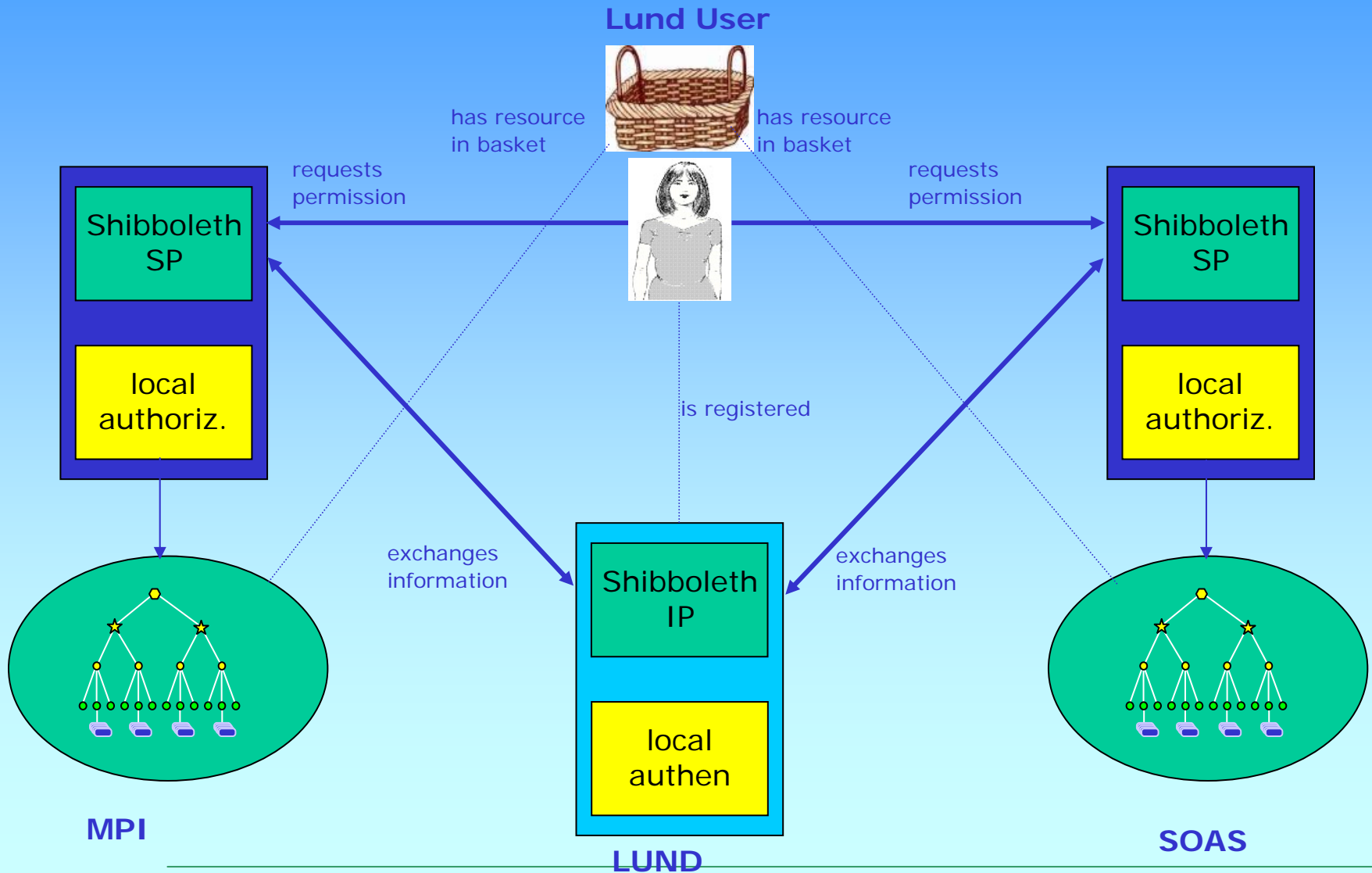
Part of the solution: **Shibboleth (1.3)**

- already widely used

- promoted actively by some national bodies and funding organizations.

---

# DAM-LR Shibboleth Scenario





# User Authentication & Authorization

## Shibboleth not perfect for our domain

Shibboleth well suited for authorization by federation wide agreed groups

Managing access for individuals requires uids

Federation wide unique uid.

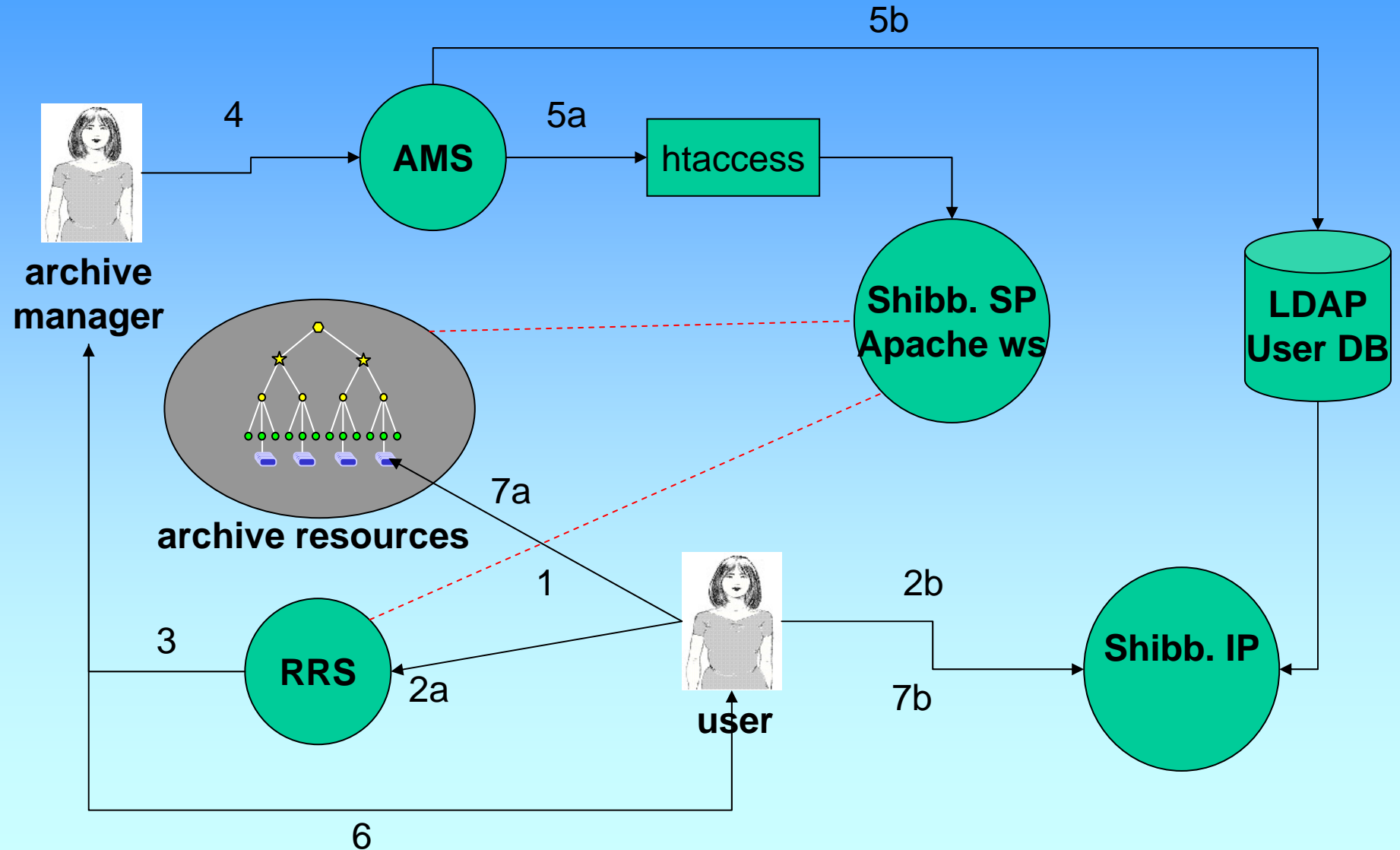
## Applications need access too!

Shibboleth depends on web-browser functionality: jscript, cookies,...

Applications need to mimic web-browser behavior?

---

# Authentication & Authorization Components





# DAM-LR achievements

Integrated metadata domain is completed

HS installations have been set up.

Partial site replication

Current Shibboleth test installations have been finalized and moved to the production servers.

---

# DAM-LR unfinished business

## Regional archives:

Not all have good back-up facilities/high network band width

Synchronization process becomes important (to archives with more secure storage)

Required: controlled synchronization

## Legal issues:

Depositors rights/restrictive conditions for deposits

Content of resources – vulnerability of archive holders

Fuzzy boundary between depositor and user with new applications like ADDIT

---

# DAM-LR, what's next? CLARIN

CLARIN is committed to establish an integrated and interoperable **research infrastructure** of language resources and its technology. It aims at lifting the current fragmentation, offering a stable, persistent, accessible and extendable infrastructure and therefore enabling **eHumanities**.

**Integrated domain:** the resource and service centers are connected via Federation technology and form a virtually integrated domain

**Interoperable:** the resources and services will be based on Semantic Web technologies to overcome format, structure and terminological differences

**Stable:** the resources and services are offered with a high availability

**Persistent:** the resources and services are planned to be accessible for many years so that researchers can rely on them

**Accessible:** the resources and services are accessible via the web; different access methods and training possibilities are offered tailored to the needs of the communities making use of them

**Extendable:** the infrastructure is open so that new resources and services can be added easily

---