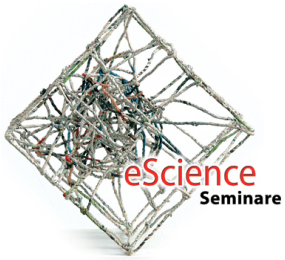
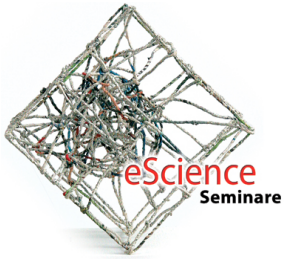


LAMUS/LAT Repository System

Daan Broeder

MPI for Psycholinguistics

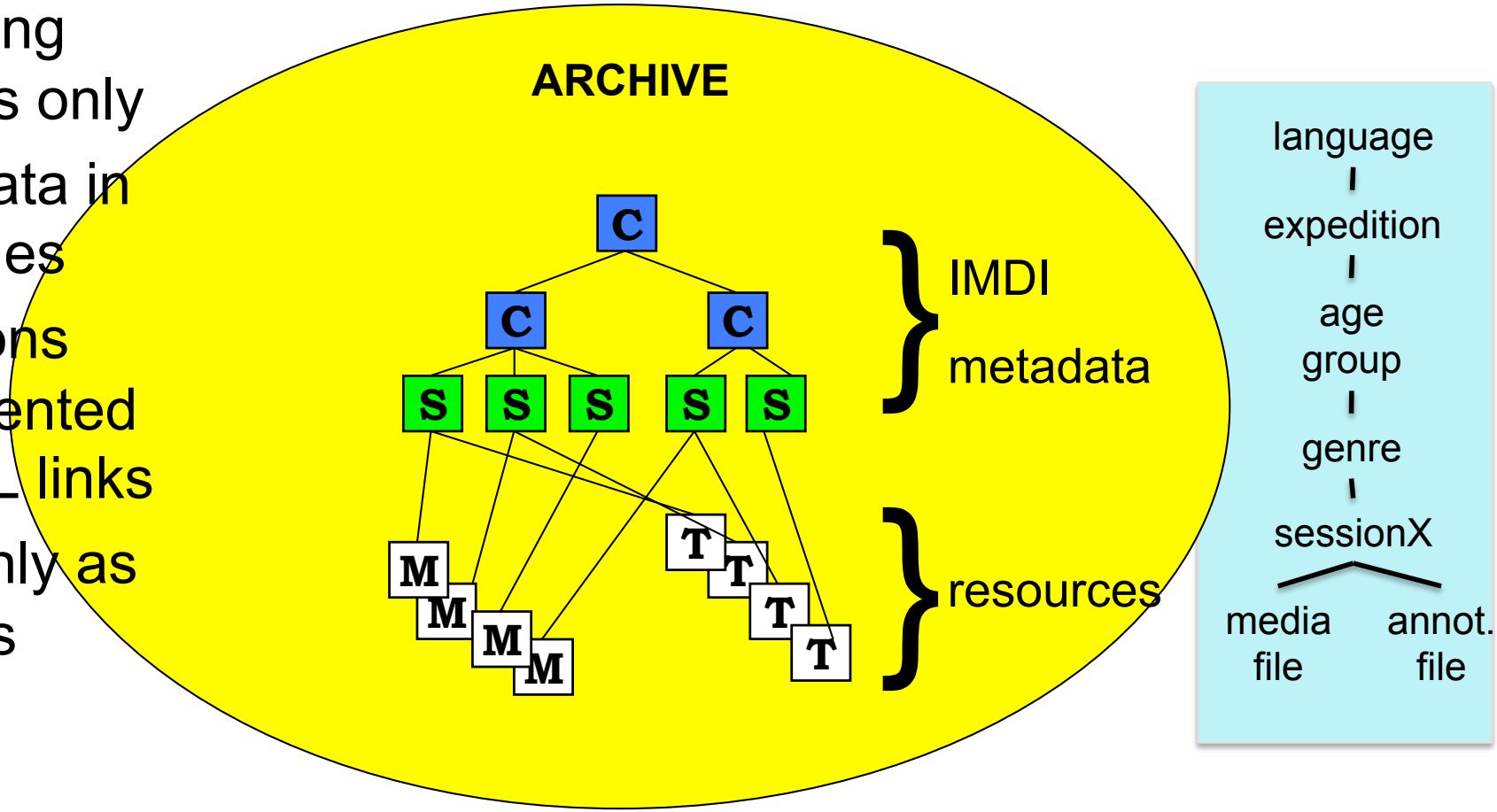


History

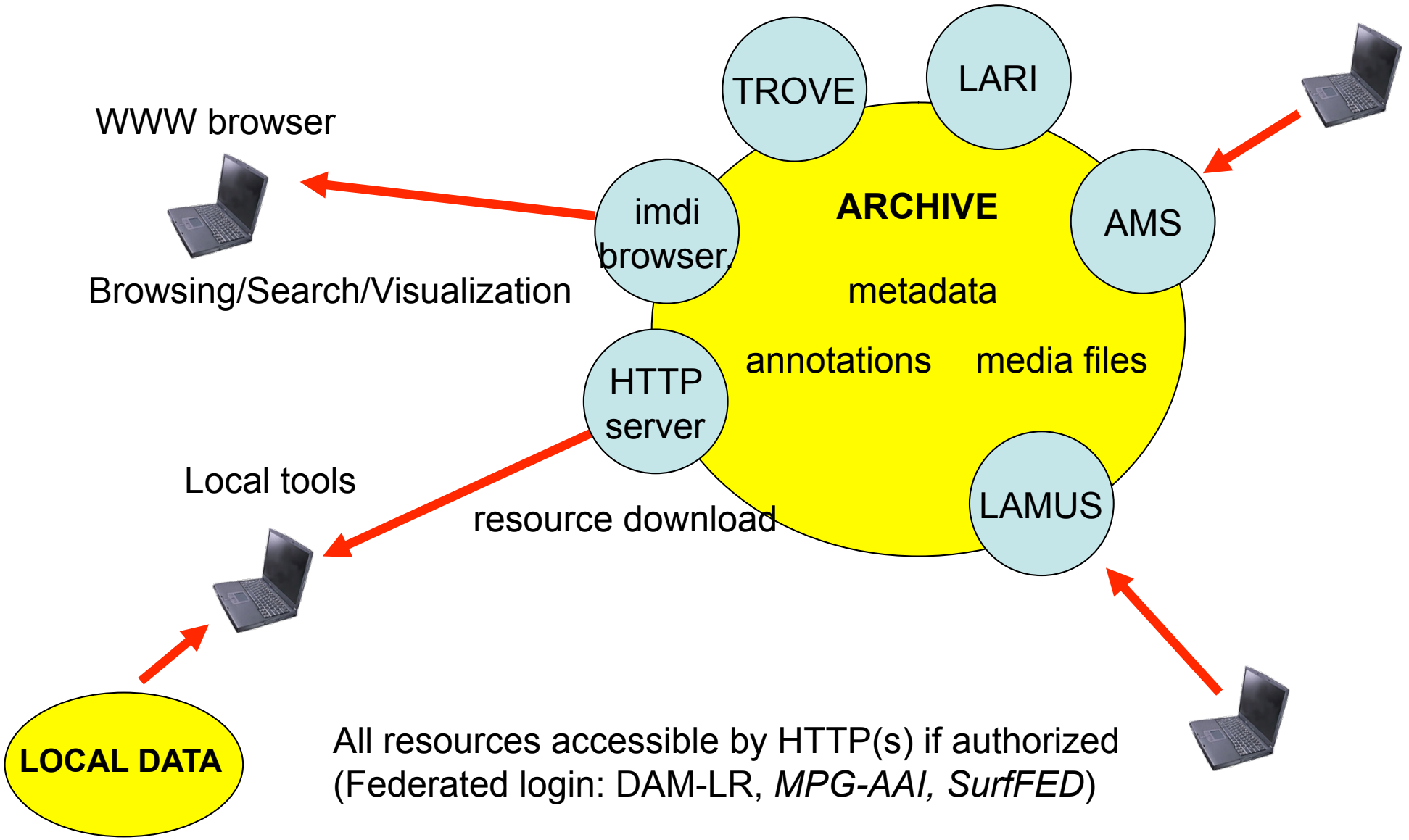
- Development started in 2000 to try solve the data organization problems at the MPI for Psyl.
- Closely linked to the IMDI metadata set for Language Resources, developed around the same time
- First version “Browsable Corpus” was basically a file-system with metadata descriptions and resource files
- Tools operating directly on the files
- The researcher’s notebook disk was just as sophisticated

Archive Data Organization

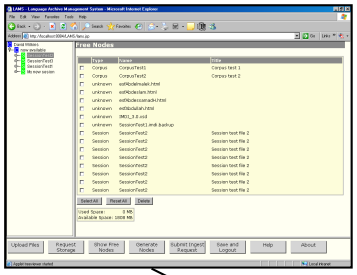
- Archiving formats only
- Metadata in XML files
- Relations represented by URL links
- DBs only as helpers

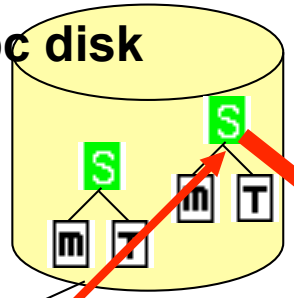


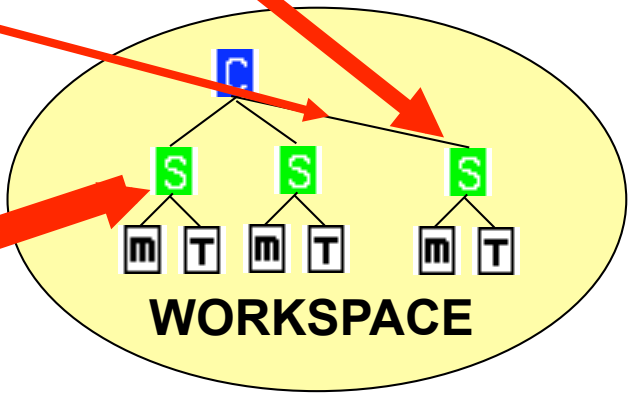
Archive Web Applications

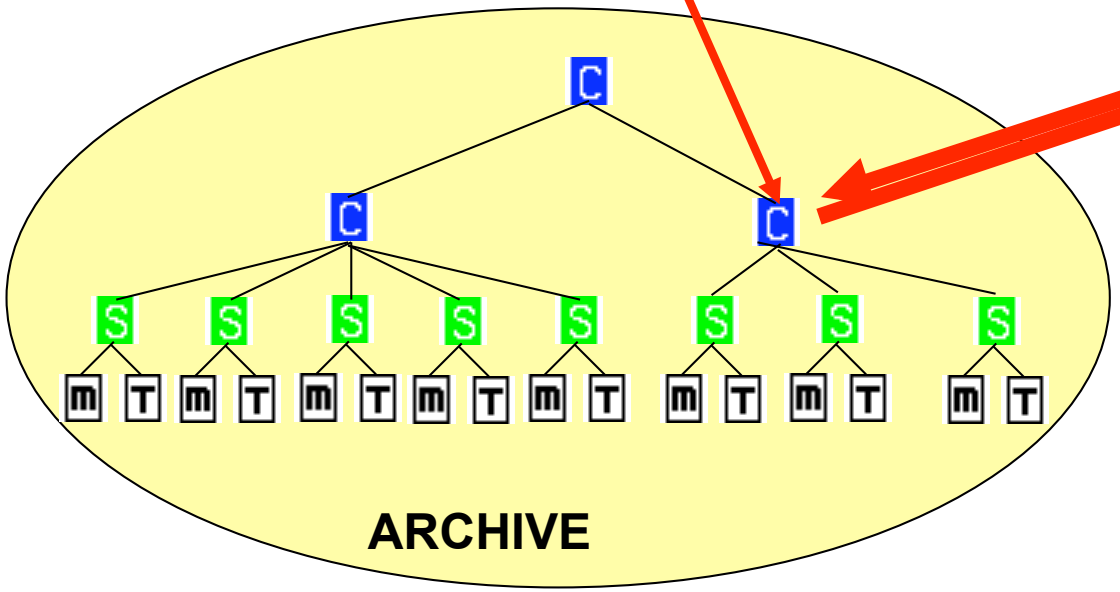


LAMUS deposition anim.

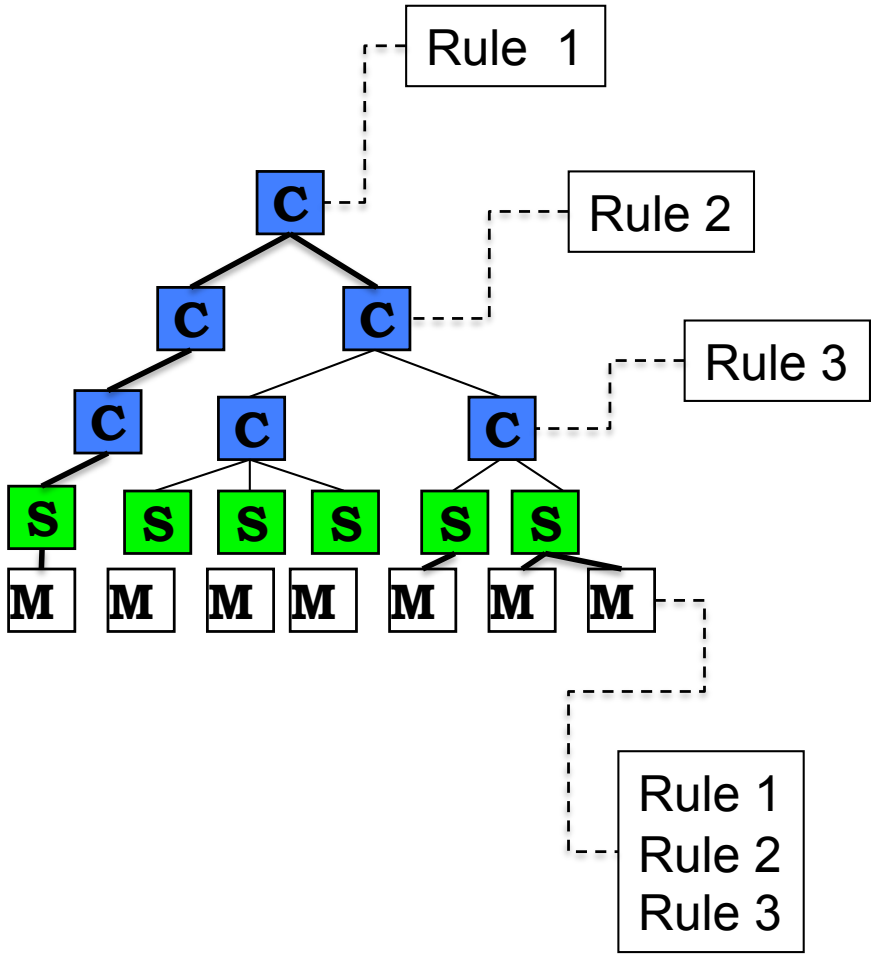






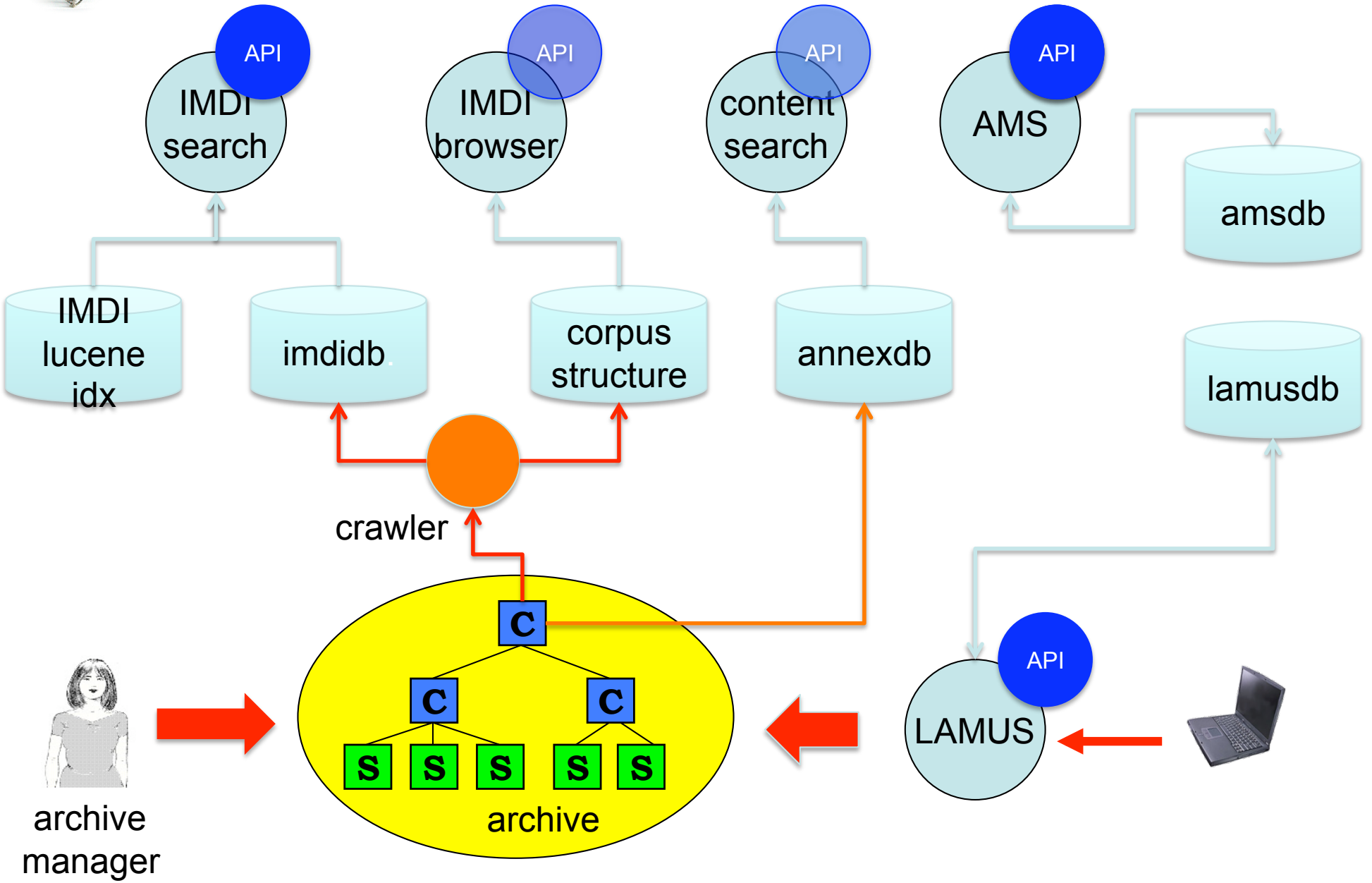


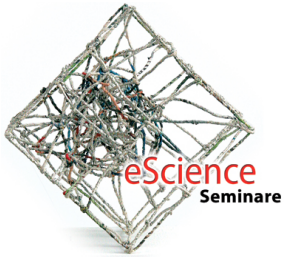
AMS - Access Management System



- User role administration: domain manager, editor
- Manage licenses & code-of-conducts
- Set access rules per media type
- Inheritance of rules, but counter rules can be set too.

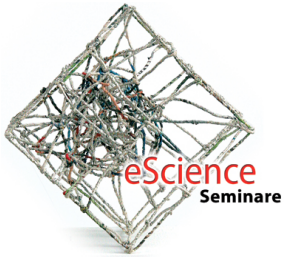
Archive Access & Administration





Versioning

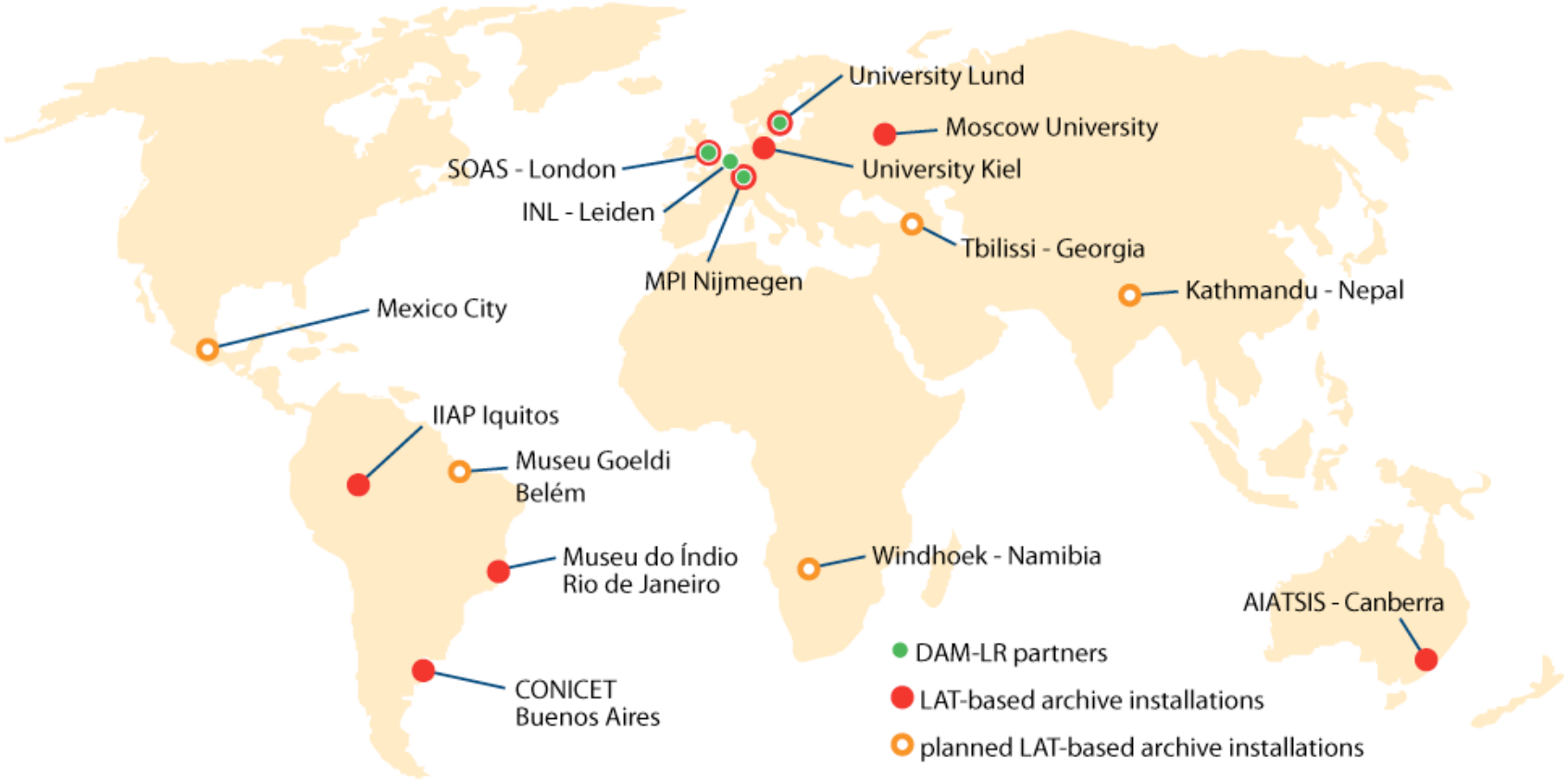
- Never throw “anything” away
  - Somebody may have linked to it.
  - Save “deleted” & “replaced” objects in special version archive
- Versioning policy
  - Need explicit “replace” from user
  - Alowed for resources and session files
  - New identifier is issued to the new version
  - Version relation is created between “old” and “new” version
  - Unsolved problem: modify a resource -> modify all containing collections



Persistent Identifiers

- LAMUS/LAT uses the Handle System as a PID framework
- Every new object gets a PID when ingested in the archive.
- Not built into LAMUS but is a configurable option
  - Still costs (some) money
  - Expresses a commitment that not all can make

LAMUS/LAT Installations



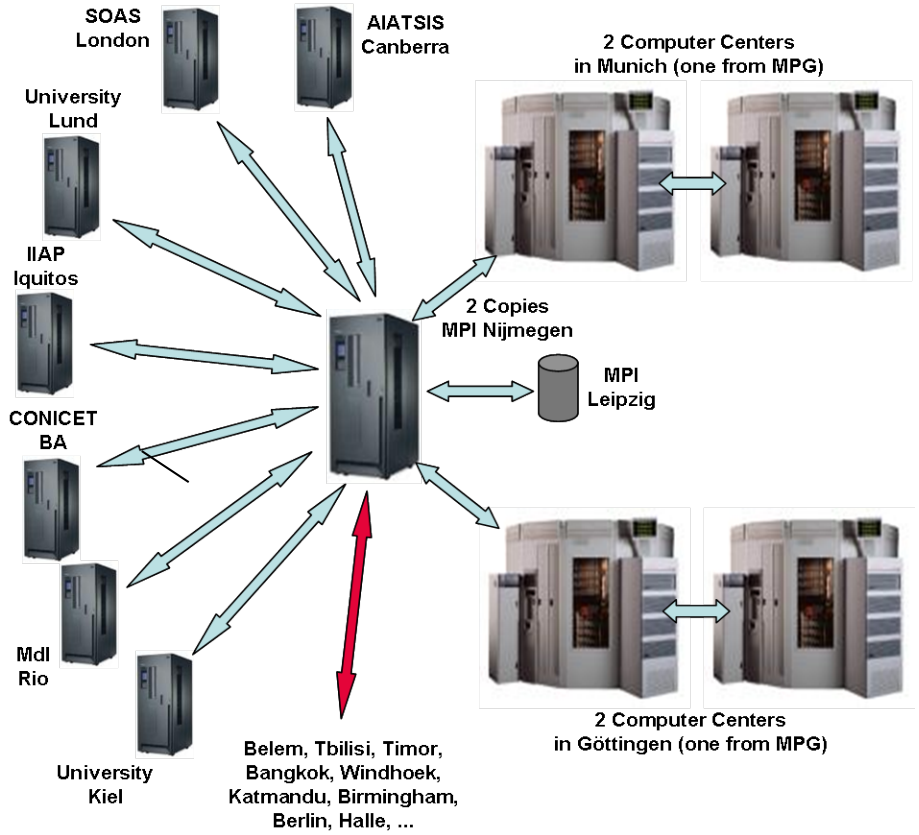
Data Synchronization I

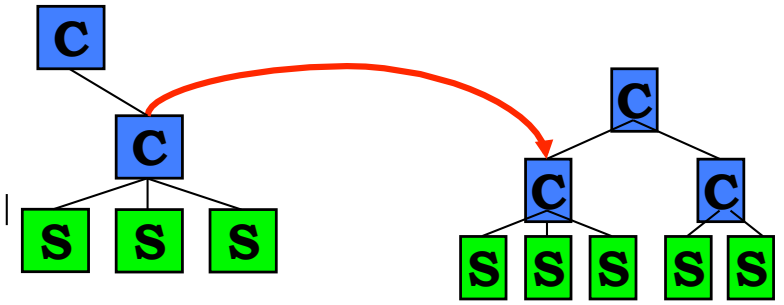
Synchronization physical structure

- Use “rsync” tool
- Complete replication
- No special conditions possible
- Use for backup to computing centers

Synchronization logical structure

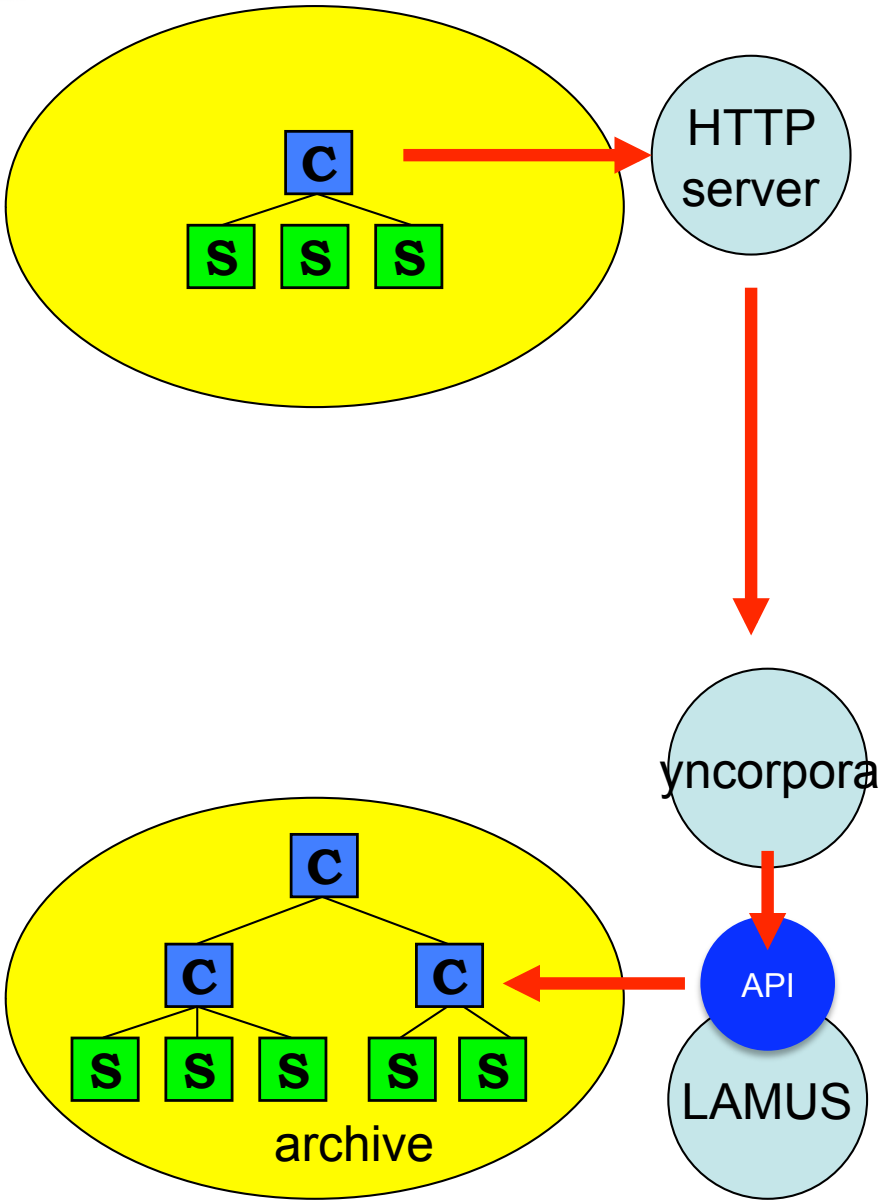
- *Special software needed*
- Per corpus copy to a selected target
- Owner can make special exceptions
- Use to synchronize between archives with different content





Logical synchronization

Data Synchronization II

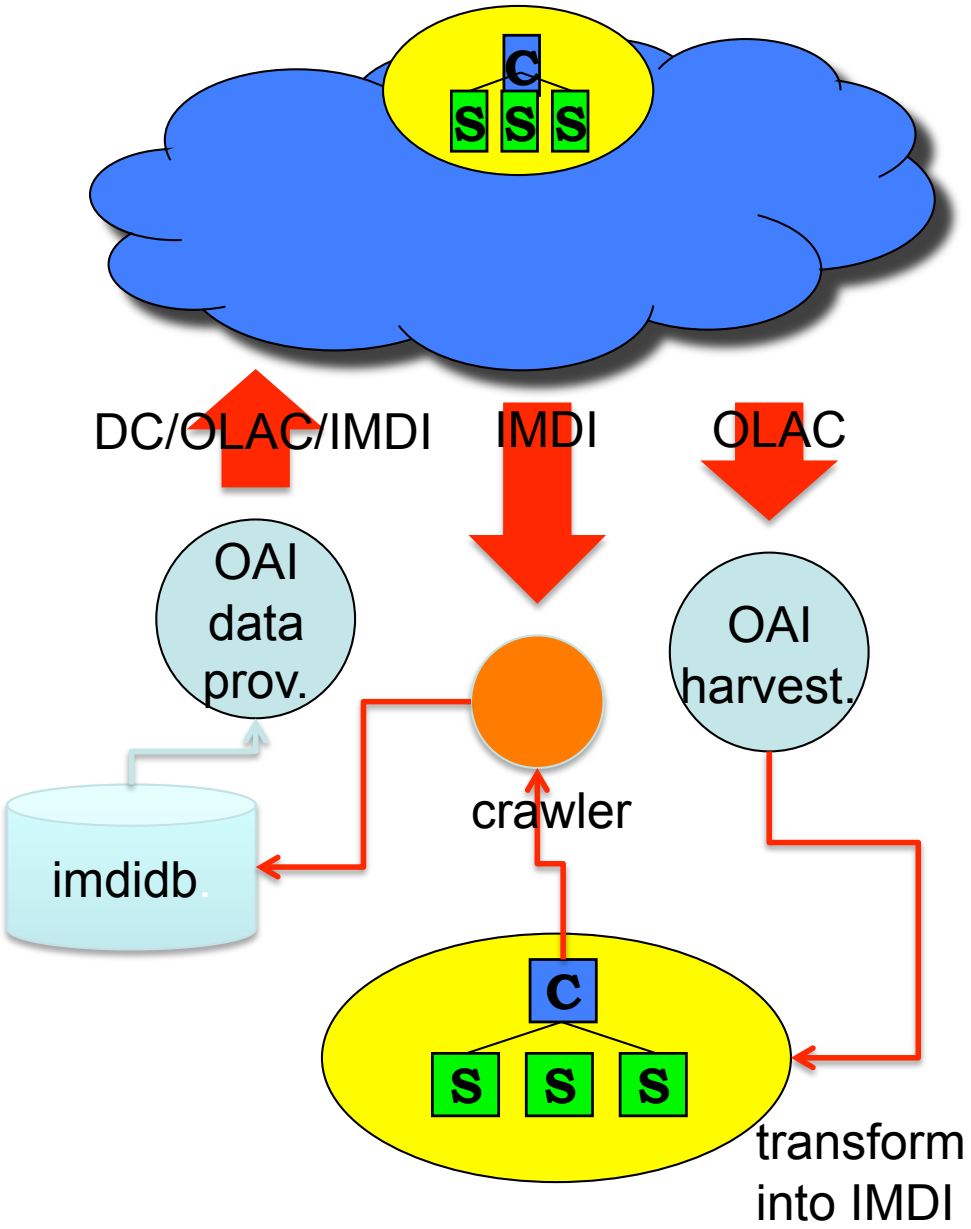


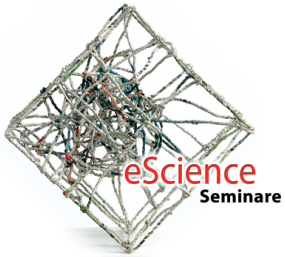
yncorpora: complex logic to compare corpus trees and determine

- what is new
- what to replace
- what to add
- what to delete

Metadata Interoperability

- Harvest IMDI metadata from other archives: Unified IMDI catalog
- Offer metadata to others: DC, OLAC, IMDI format
- Harvest metadata from other archives using OAI protocol.
- All this to arrive at a unified catalog for IMDI & OLAC metadata

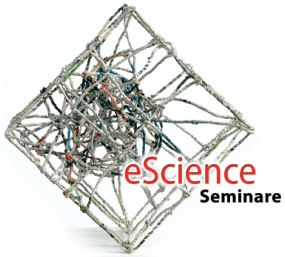




Status & Costs

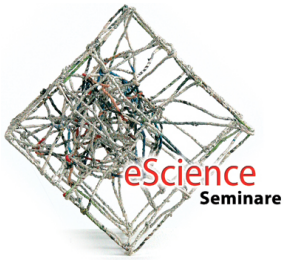
LAMUS/LAT is a dedicated system for managing & utilizing language resources

- Currently 80000 sessions, 300000 digital resources, 20 TB of data
- LAMUS/LAT system now installed at about 8 institutes
- repository system incl. metadata, access management etc: ±100.000 lines of code
- utilization software much more heterogeneous: ±230.000 lines of code
- creation costs of repository system: ±1 M€
- software maintenance costs for repository system ±60 k€/ year



Current & Future Challenges

- Further implement web services (REST, XML-RPC & SOAP) for services: AMS, content search, …
- Installation package
- Code refactoring to increase stability & maintainability
- Interaction with other Repository Systems and other information sources: Inter & Harve
- Adapt to new types of data and metadata
  - Neuro imaging data
  - CLARIN metadata infrastructure



The End

Thank you for your attention