Managing very large Multimedia Archives and their Integration into Federations

Daan Broeder, Eric Auer, Marc Kemps-Snijders, Han Sloetjes, Peter Wittenburg, Claus Zinn

Max-Planck Institute for Psycholinguistics

2008 VLDL workshop, Aarhus

Content

- The MPI Archive and its collections
- Data organization model
- Archive interoperability projects & technologies
- Future developments

Nijmegen Language Archive

- MPI for Psyl. Corpora: Child language, bilingualism, gesture, sign language, Corpus Spoken Dutch, acquisition corpora, etc.

- Archive for the DOBES project: Endangered Language Documentation resources
  - Representative record of a language in its cultural context
  - May help in maintaining and revitalizing languages

- Hosting and inviting corpora from other projects in need, (even not strictly linguistic material)
  - DBD, NGT, Eibl Eibesfeldt human ethol. collection, …
- Maintain metadata catalogue for IMDI described resources
  - BAS, C-ORAL-ROM, …

Mostly annotated audio/video recordings

30 Terabyte, 53.000 AV resources, 24.000 annotation files,

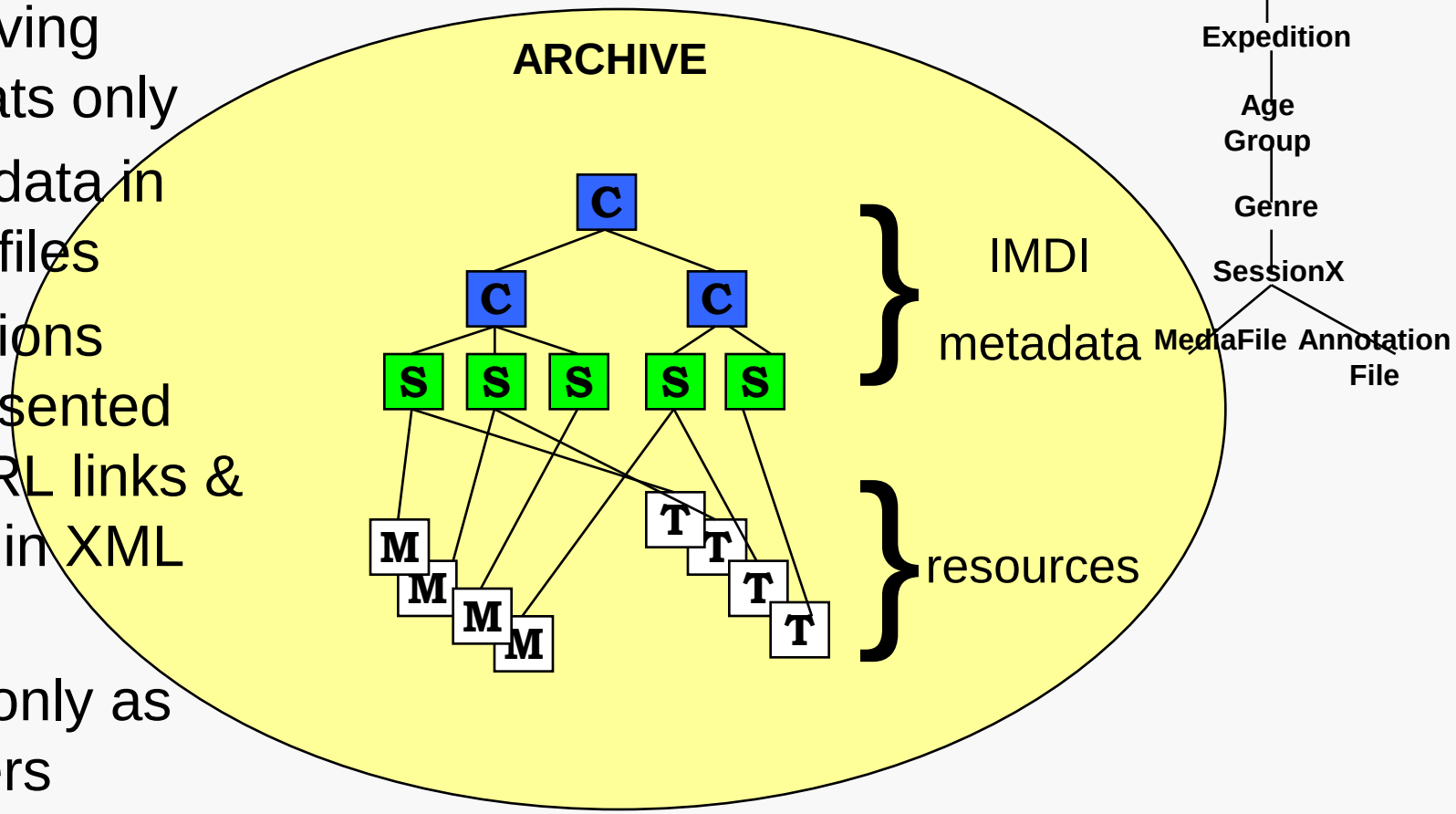
60 Mio annotations, lexicons, sketch grammars, etc.

Archive Management

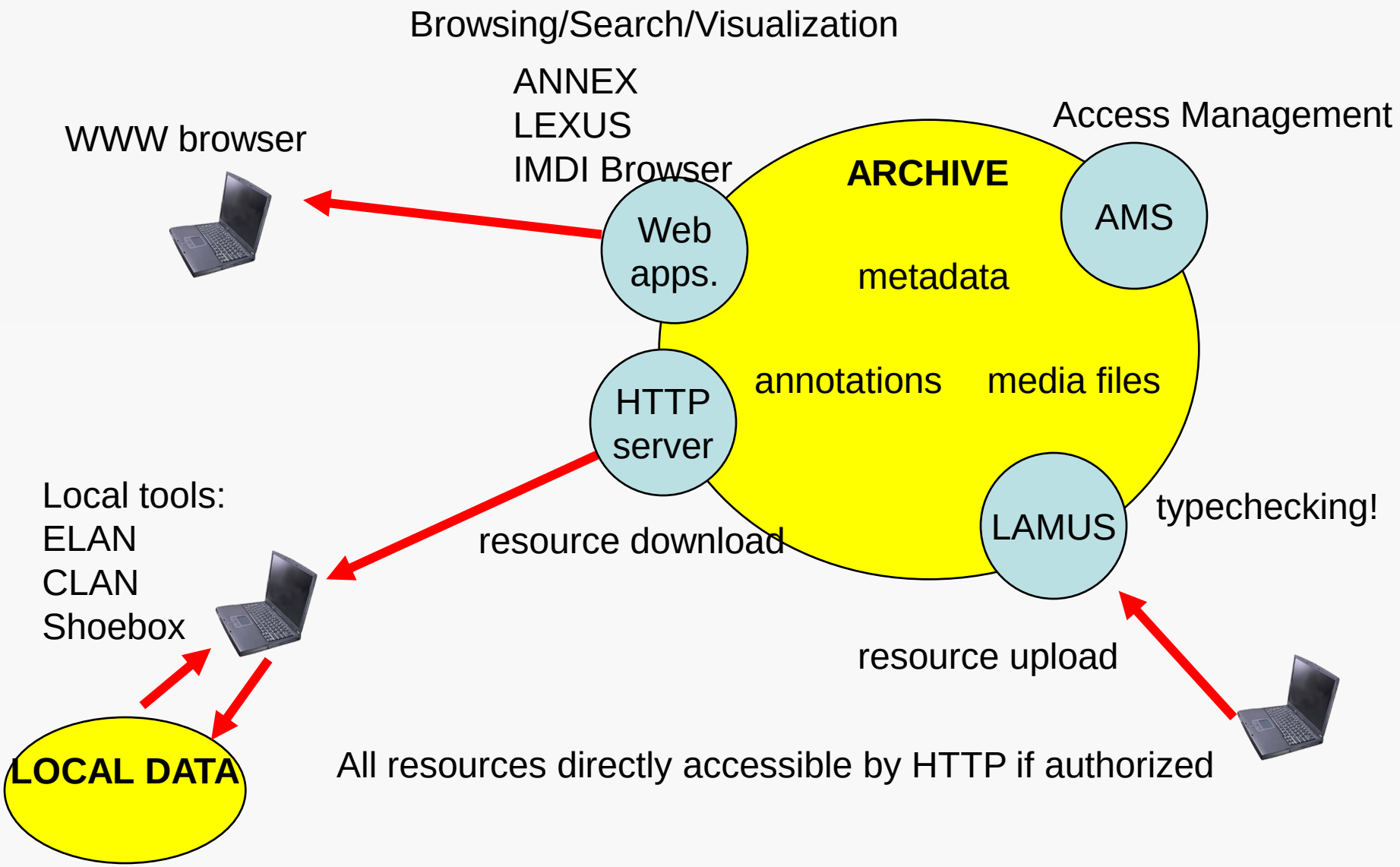
- We are an archive, preservation is our first concern but usage is important and providing this takes up most resources.
- Management not (only) a question of the amount of data although its is important for:
  - Making safe copies
  - Managing storage technology change
- Organization of the data
  - Describing & labeling the data – metadata
  - Allowing user access to the data
    - Access rights configurable for every individual resource
  - Live Archive so allow depositors to
    - Upload data into the archive
    - Provide new versions of existing resources
    - Add new information & comments for existing resources

Archive Data Organization

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



Archive Access



Language Archiving Technology

preparation

ELAN/LEXUS/SYNPATHY
Annotation + Lexicon

Shoebox/CHAT
Transcriber
XML

integration

IMDI
Data Organization, Metadata

LAMUS
Data Uploading and Management
Access Management

Archive Grid
Federation

Data Archiving and Copying

utilization

IMDI / GIS
Metadata Browsing & Searching

ANNEX/LEXUS/IMEX/
TROVA
Complex Access via Web

ADDIT/VICOS/MEL
Enrichments/Views

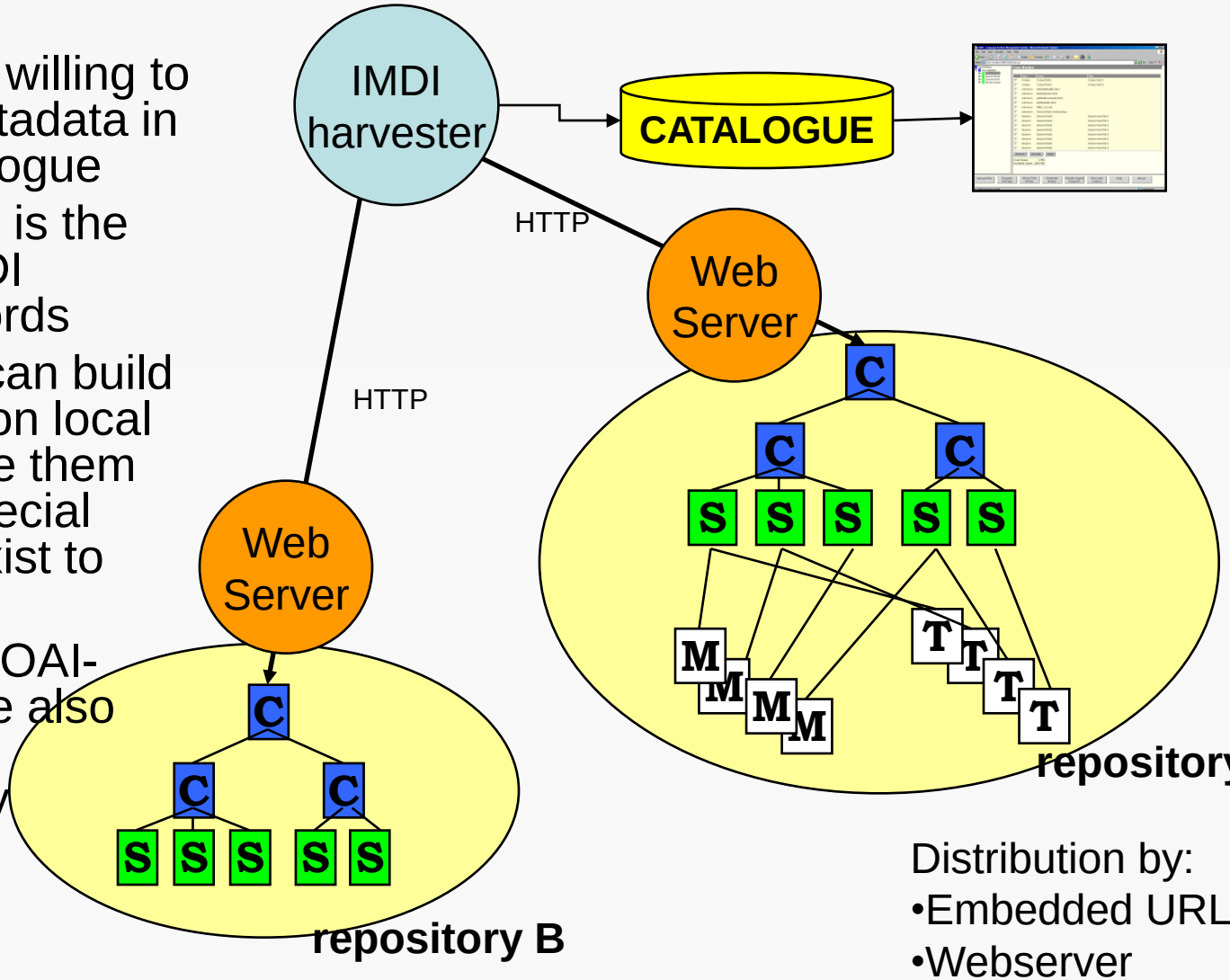
ODIT/ISOcat
Ontology
management
framework

LAT to support operations during resource life-time

support standards where possible

Distributed Repositories

- Organizations willing to show their metadata in a central catalogue
- Only condition is the offering of IMDI metadata records
- Researchers can build IMDI corpora on local disks and have them harvested. Special client apps. exist to support this.
- Different from OAI-PMH which we also support for interoperability



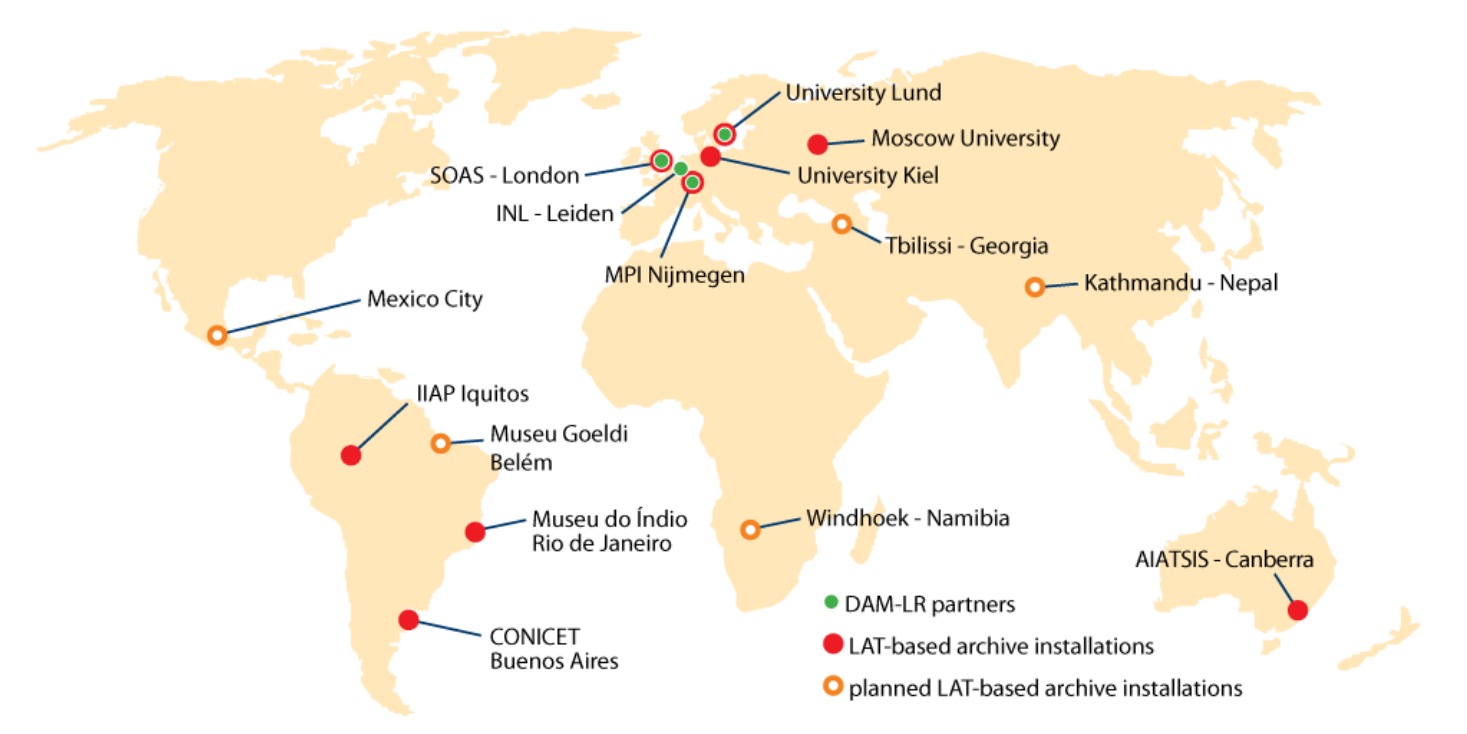
Distribution by:
- Embedded URL link
- Webserver
- Low tech!!!

DoBeS project (2000-…)

(funded by the Volkswagenstiftung)

40 language teams from the DOBES program documenting about 60 languages and working independently

Regional Archives Initiative



Cooperation of MPI with other organizations interested in EL

Receive Installations of the MPI/LAT archiving software

- Encourage local resource collecting & archiving
- Foster local responsibility for resources

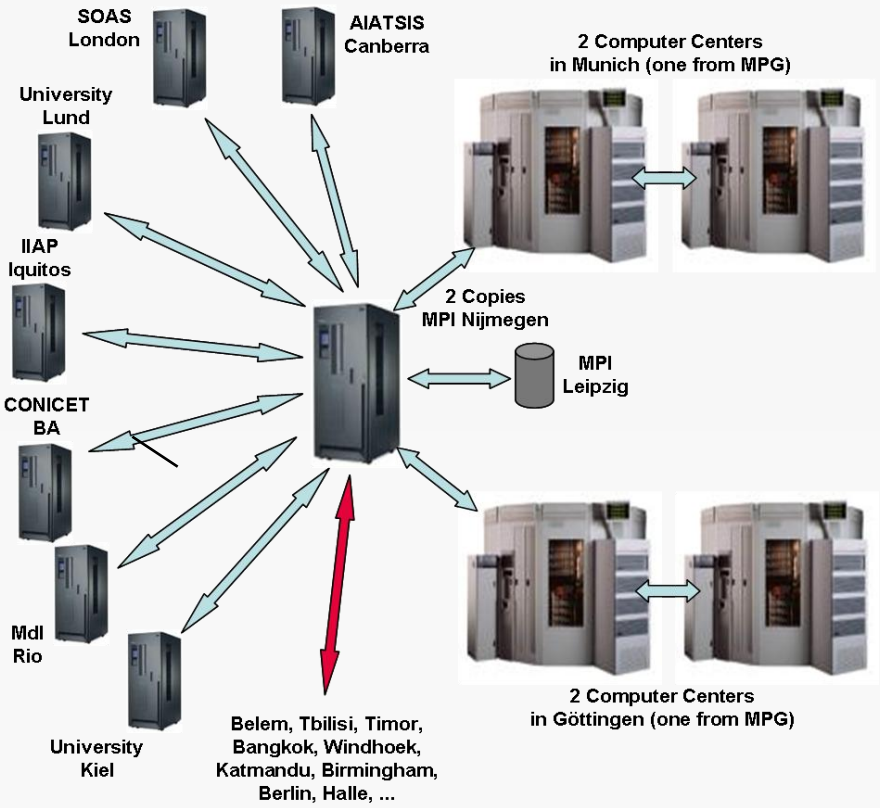
Data Synchronization

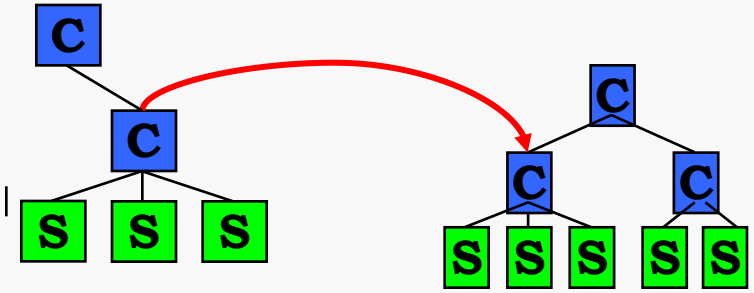
Data sync. physical structure

- Use “rsync” software
- Complete replication
- No special conditions possible
- Use for backup to comp. centers

Data sync. logical structure

- *Special software needed*
- Per corpus copy to a selected target
- Owner can make special exemptions
- Use to sync between archives

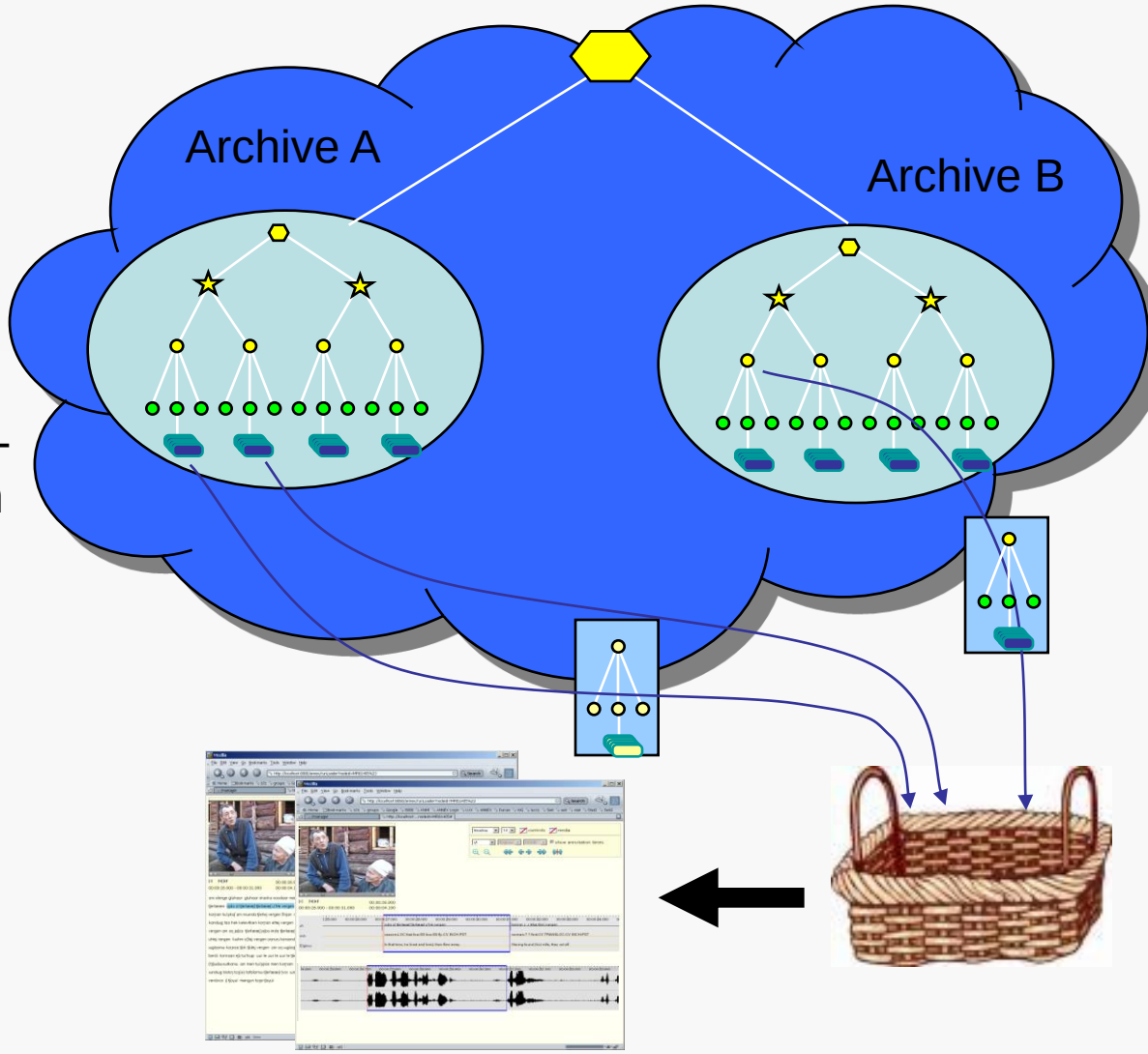




Logical sync.

Why federate ?

- Allow researchers to build virtual collections
- Requires interoperability different levels
  - Authentication & authorization
  - Selection of resources – single metadata domain
  - Unified way of referring to resources.
  - Format interoperability
  - Semantic interoperability



DAM-LR EU project (2005-2007)

(Small) EU project on archive integration of 4 partners corpus/computational linguistics and endangered language documentation

- Resource discovery: sharing a single metadata set for searching & browsing
- AAI: single user identity, single sign-on.
- Referencing and citing “archived resources” using a single persistent identifier system with added services

AAI with Shibboleth

- Successfully installed 3 IdPs and SPs sets
- Tried to invent own attribute set, but eduPerson should be sufficient.
- Managing authorization with Shibboleth is not perfect for our domain
  - Shibboleth well suited for authorization by federation wide agreed groups
  - Managing access for individuals requires federation wide unique uid.
  - The SP should have a record for every user they grant access to
- Applications need access too!

Persistent Identifier Framework

Avoid dead links by separating resource name and location using a resolving service to translate the name into a URL.

- DAM-LR opted for the Handle System (HS) (also the basis for DOI)
  - Robust, scalable, secure, multiple URL support, well used
- Every partner runs own resolving service with a backup for the other partners.
- HS optional component in LAT archiving software.
  - Not every repository can make the commitment
- Own services build on top of HS
  - Distribution of authorization information for resource copies
  - *Many more services are possible*
- HS problems:
  - Missing part identifiers like in ARK
  - Problems with standardization, W3C only likes URIs

Future projects: CLARIN

Common Language Resources and Technology Infrastructure

- Much larger then DAM-LR
- Will (probably) adopt:
  - HS as a PID framework
    - Develop some extra services
  - Shibboleth for AAI
    - Find solution for application authentication
- Metadata framework must be much more flexible
  - Considering a Component Framework much like Application Profiles.
  - Semantic interoperability using ISO DatCat

The End

Thank you for your kind attention