Metadata Catalogue Issues

Daan Broeder

Max-Planck Institute for Psycholinguistics

DAM-LR

Introduction

- Methods of registering resources
- Metadata
- Making metadata interoperable
- Exposing metadata
- Facilitating resource discovery

Methods or Registring Resources

- By using content
  - Google a.o. for text-type files
- By using (structured) metadata: Metadata catalogue
  - Using either a very simple MD set like DublinCore (DC) 15 elements
  - Or more elaborate ones like TEI, IMDI, OLAC, MPEG7, …

DAM-LR

Metadata set – Dublin Core

Content	IPR	Instance
Title	Creator	Date
Subject	Publisher	Type
Description	Contributor	Format
Language	Rights	Identifier
Relation		
Coverage		
Source		

DC Set

DC Example

DC.Title = “Building with Lego”

DC.Title /Alternative = “Lego series part I”

DC.Creator = “L. Smith”

DC.Subject /LCSH = “Building”

DC.Description/Abstract = “…….”

DC.Language/ ISO639-2 = “eng”

DAM-LR

Metadata set – DC example

<dc:identifier>oai:www.mpi.nl:1839/00-0000-0000-0000-0009-E</dc:identifier>
<dc:title>SI011020</dc:title>
<dc:date>1972-03-26</dc:date>
<dc:description>
in der Kueche beim Fruehstueck, SIM moechte Kuchen haben
</dc:description>
<dc:coverage>Germany</dc:coverage>
<dc:publisher>
Corpus Manager - Max-Planck Institute for Psycholinguistics
</dc:publisher>
<dc:description>
in der Kueche beim Fruehstueck, SIM moechte Kuchen haben
</dc:description>
<dc:language>RFC1766:x-sil-GER</dc:language>
<dc:contributor>Annotator</dc:contributor>
<dc:format>text/x-chat</dc:format>

….

DAM-LR

Metadata set – IMDI example

<?XML … >

<METATRANSCRIPT DATE="2005-11-28" FormatId="IMDI 3.0" ….>

<Session><Name>Hamadi siblings</Name><Title> Hamadi siblings DBD_ARY_02_21_03_012 excerpt</Title>

<MDGroup>

<Location><Continent>Europe</Continent>

<Country>Netherlands</Country>

<Region>Goirle, Brabant</Region></Location>….

<Content>

<Genre>Discourse</Genre><SubGenre>Conversation</SubGenre>….

</Content>

<Actors>

<Actor><Name>Abdelkrim Hamadi</Name>.

<Language>Dutch<Language><Language>Arabic, Moroccan Spoken</Language></Actor>

<Actor>…<Actor></Actors>

<Resources>…</Resources>

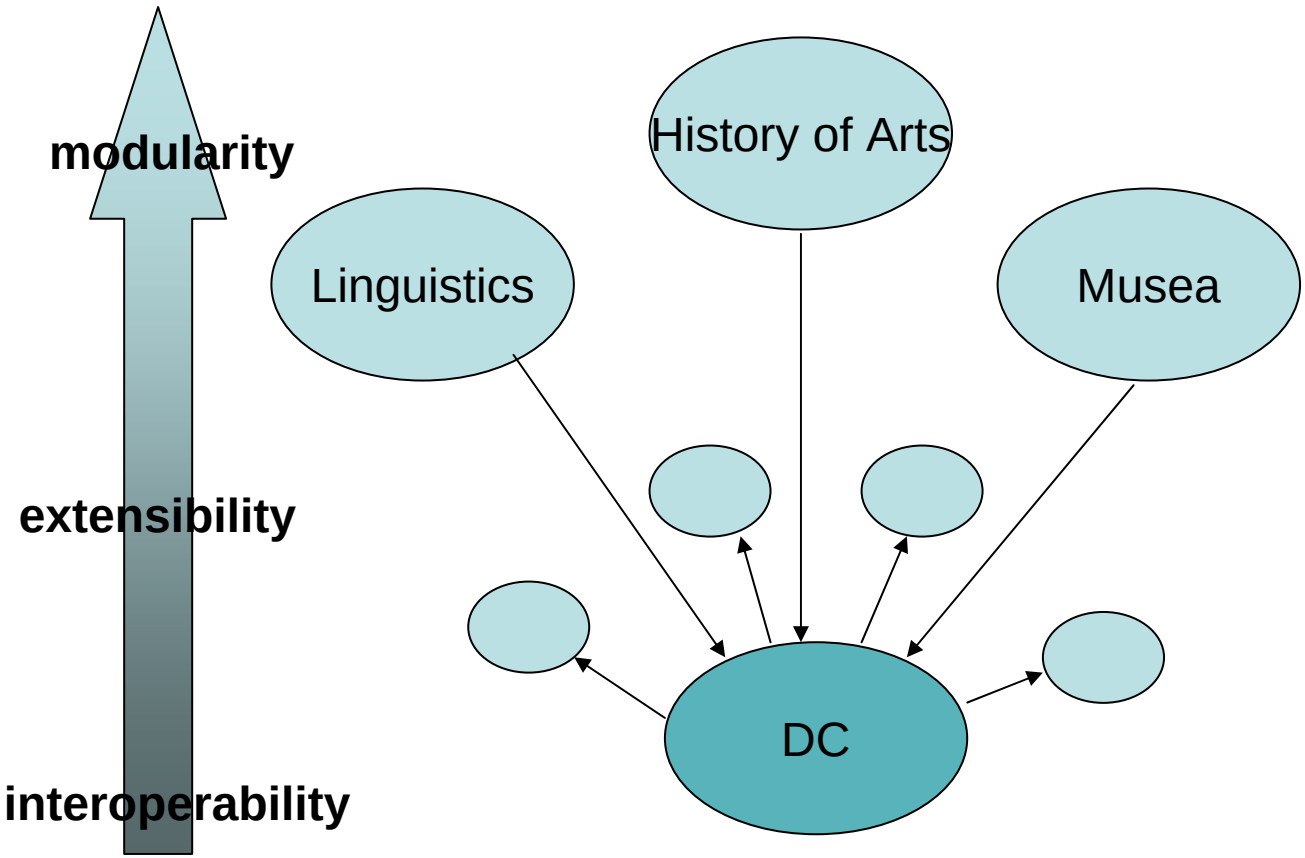
….

DAM-LR

Metadata pidginization & creolization

Conflicts between

- Simplicity for interoperability
- Expressiveness for complete description



DAM-LR

Metadata Interoperability

Goals:

- Metadata search/browsing across repository boundaries as if there is one virtual domain.
- Appropriate terminology for users of all repositories when specifying a MD query or viewing MD records

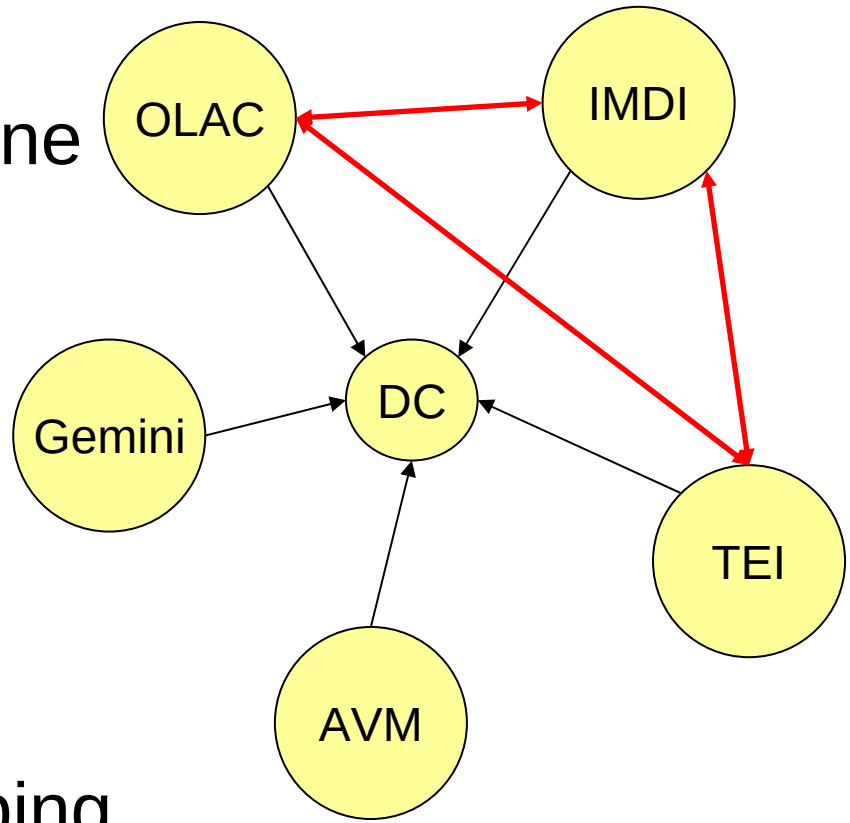
Possible solutions:

- All repositories use the same metadata set
  - Not very practical. For sufficiently accurate description different (sub-)domains need different sets.
  - There are many existing metadata repositories with own sets & procedures that are difficult to change.
  - Already difficult to use one set only in a single repository
- Use a single set for exchange (->mapping by producer required)
- Use an interoperable metadata framework

DAM-LR

Metadata Mapping

- Metadata crosswalk
  - Effective if same discipline
  - Many mapping rules
- Switching across, use pivot set.
  - Information loss
  - Limited number of mapping rules



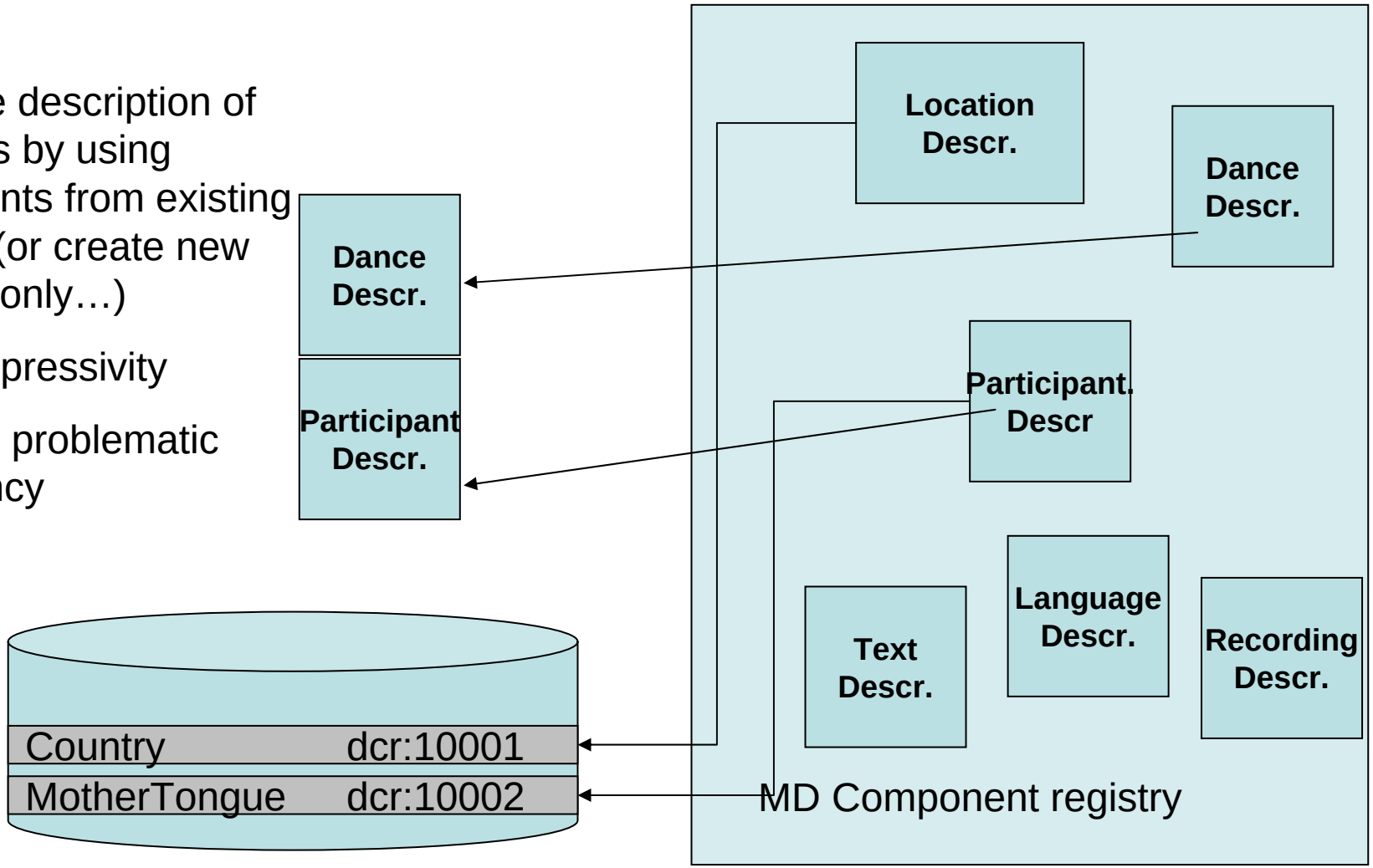
Astronomy Visualization Metadata (AVM)

UK geo-spatial metadata (Gemini)

Metadata Framework: a component approach

Adequate description of resources by using components from existing MD sets (or create new ones but only…)

- Good expressivity
- Possible problematic consistency



ISO standard DCR: concept registry

DAM-LR

Metadata Exposure

- OAI - Open Archives Initiative
- IMDI Framework
- Repository Systems: Fedora, DSpace, … (OAI + METS built-in)
- Data Grid solutions: SRB, …

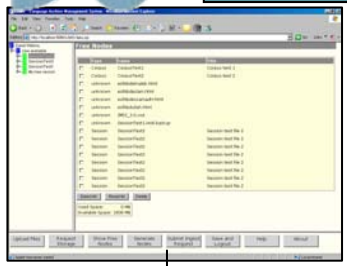
DAM-LR

Open Archives Initiative – OAI

OAI-PMH (protocol for metadata harvesting)

- low-barrier interoperability framework for archives
- Originally dev. for e-print servers to expose MD for documents
- OAI data providers offer metadata for harvesting to OAI service providers
- Different metadata sets are allowed but DC MD should always be available
- Harvesting uses XML over HTTP protocol

OAI metadata distribution



Service Provider

MD Search

MD CATALOG

others

- Web-services
- REST interf.

OAI-PMH

OAI Harvester

OAI repository

MD DB

Data Provider 1

OAI repository

Data Provider 2

OAI repository

Data Provider 3

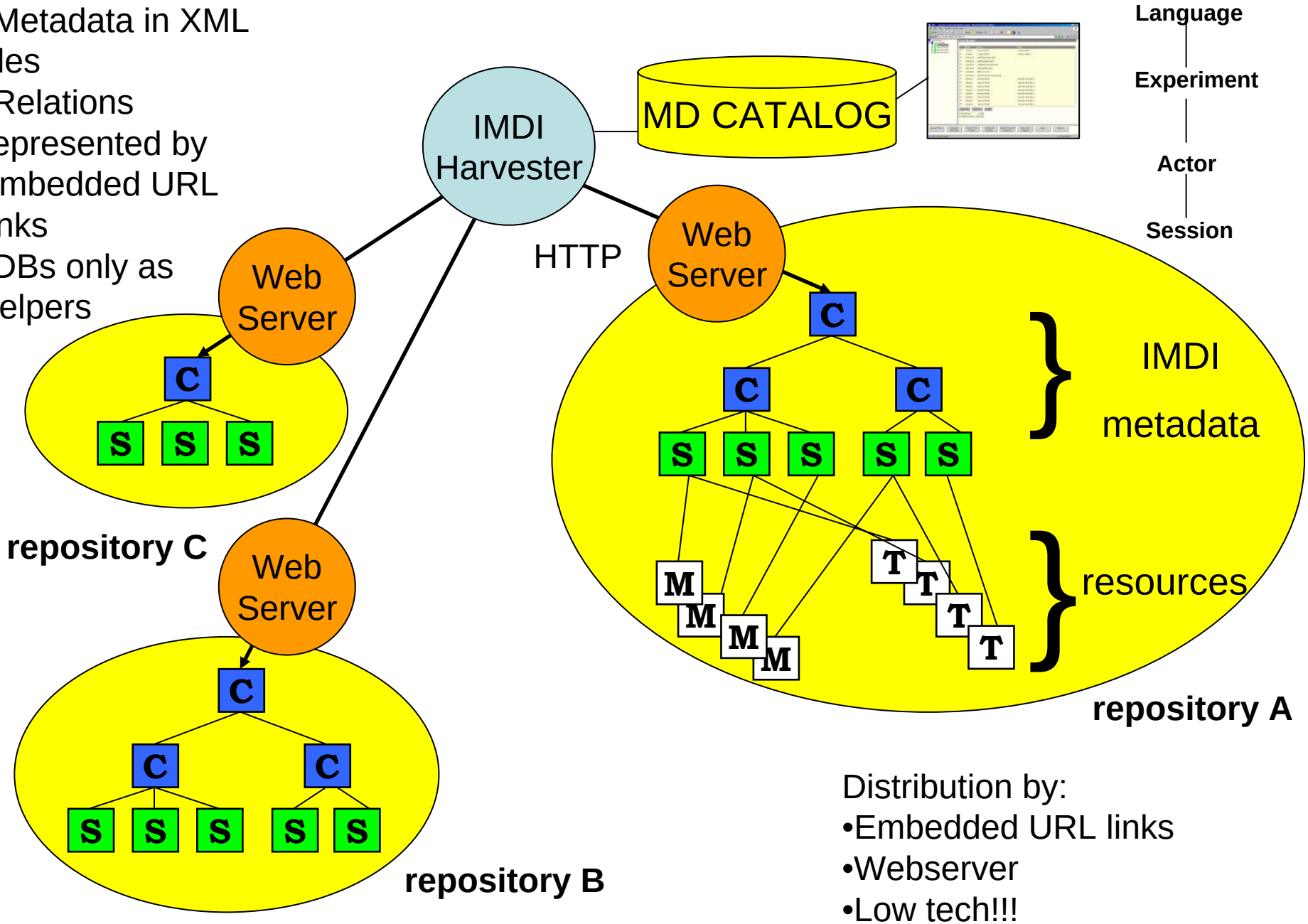
DAM-LR

OAI

- Easy to implement, but ….
- Easy to expand on: exchange within own community , OLAC metadata
- Relying only on the “required” metadata part (DC) gives information loss.
  - However communities may agree on their own MD set and see the DC only as a courtesy.
  - Or mapping may take place at the SP

IMDI metadata framework

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



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

DAM-LR

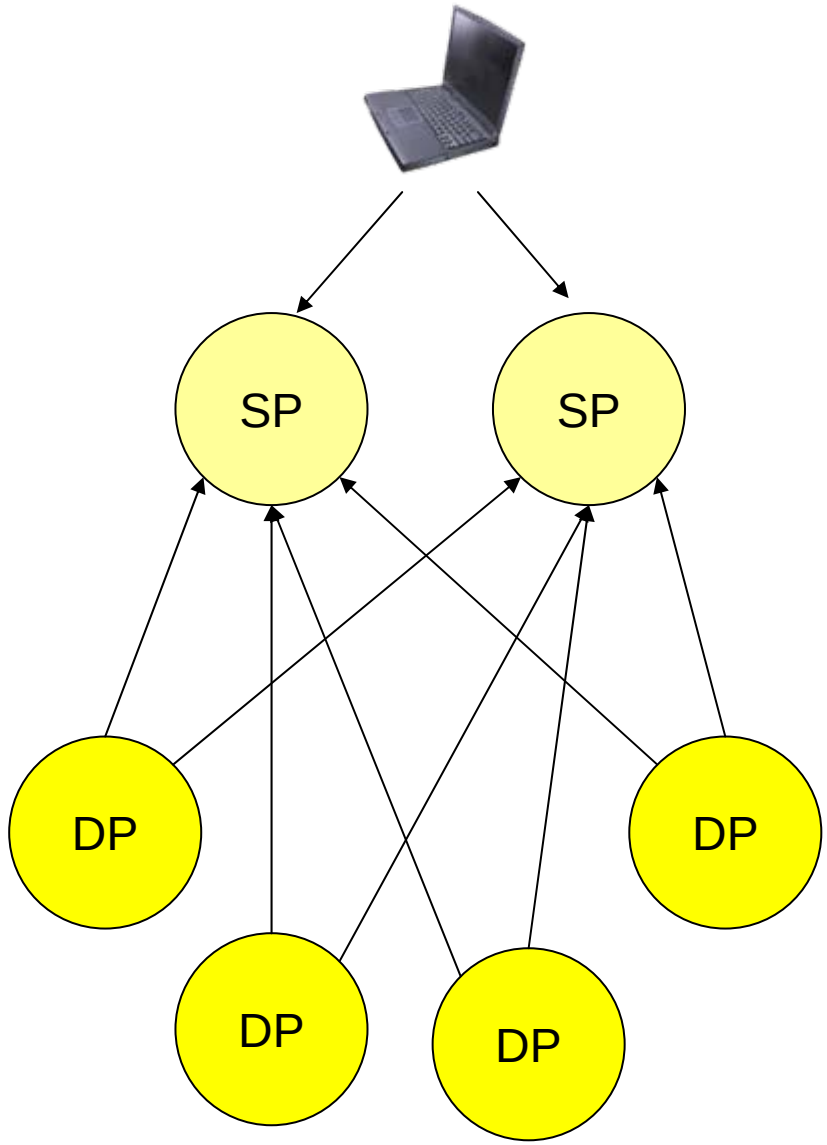
Accessing the Metadata Catalogue

- With OAI or IMDI we can expose the metadata to whoever requires it.
- How can we access it efficiently both for the service provider and the user?

DAM-LR

Complete Catalog

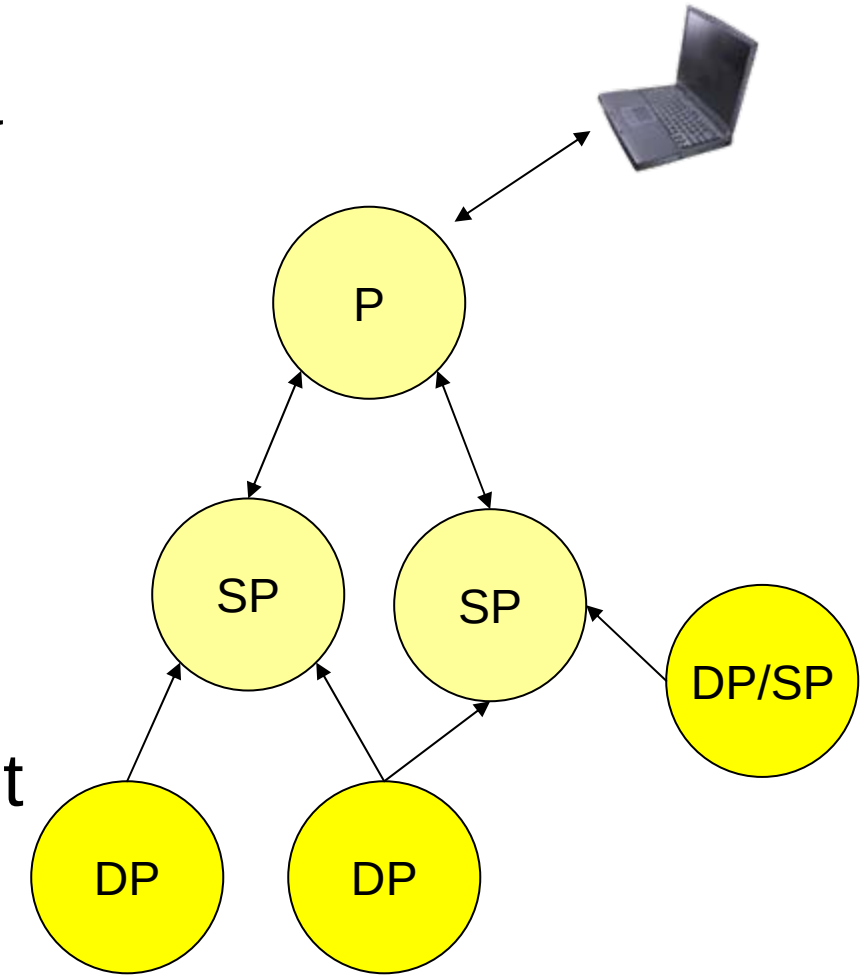
- Every service provider harvests all data providers
- Offers a MD search interface working only on its own catalog
  - “high availability” solution
  - requires many resources
- SP can map from different offered sets into its own catalogue
- Or the SP requires the DP to deliver one specific MD set.



DAM-LR

Federated Search

- A portal has no metadata catalog but broadcasts a metadata query to all appropriate SPs
- Portal can transform MD queries to appropriate specific formats used by different SPs
- Present merged result set to the user.



The End