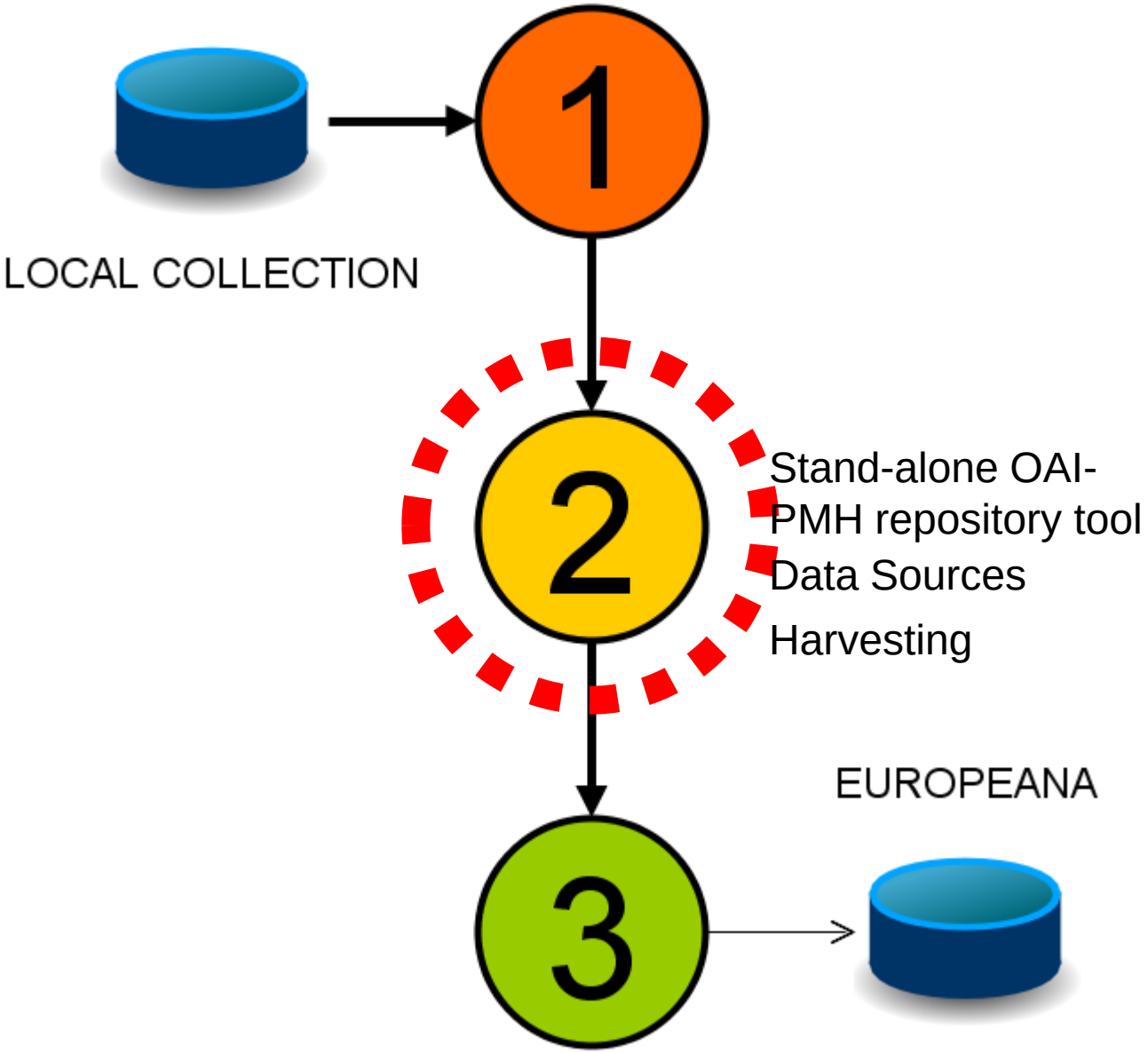
CARARE Training Workshops

Stein Runar Bergheim
Asplan Viak Internet as

europeana
carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

Stage 2: Repostiory Mgmt.



Understanding OAI-PMH

europeana

carare project

** Some of the following slides are based on : Cole , T.W., Mischo, W.H. and Habing , T.G. 2003: Introduction to the Open Archives Initiative Protocol for Metadata Harvesting, JCDL 2003, Houston, TX*

CARARE is funded by the **European Commission's ICT Policy Support Programme**

OAI Protocol for Metadata Harvesting*

Harvesting approach to interoperability at metadata level

Divides world into

- Metadata Providers
- Service Providers

Builds on HTTP, XML, & Dublin Core

europeana

carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

Data and Service Providers*

Data Providers (Repositories) refer to entities who possess resources & metadata and are willing to share metadata with others via well-defined OAI protocols

Service Providers (Harvesters) are entities who harvest metadata from Data Providers in order to supply higher-level services to users (e.g. search & discovery)

europeana

carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

Reliance on HTTP & XML*

OAI-PMH is a REpresentational State Transfer (REST) protocol (unlike RPC, SOAP)

OAI requests and responses are sent via the HTTP protocol

OAI Requests are encoded as HTTP GET or POST operations

OAI Responses are valid XML documents

europeana

carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

XML Namespaces and Schema

Consistency and data “quality” may be ensured by using XML Schema Definitions (XSD) for all responses

In this training course we are merely using OAI-PMH as a transport for arbitrary XML-formats

Schema will therefore not be dealt with in great detail

europeana

carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

The XML Response

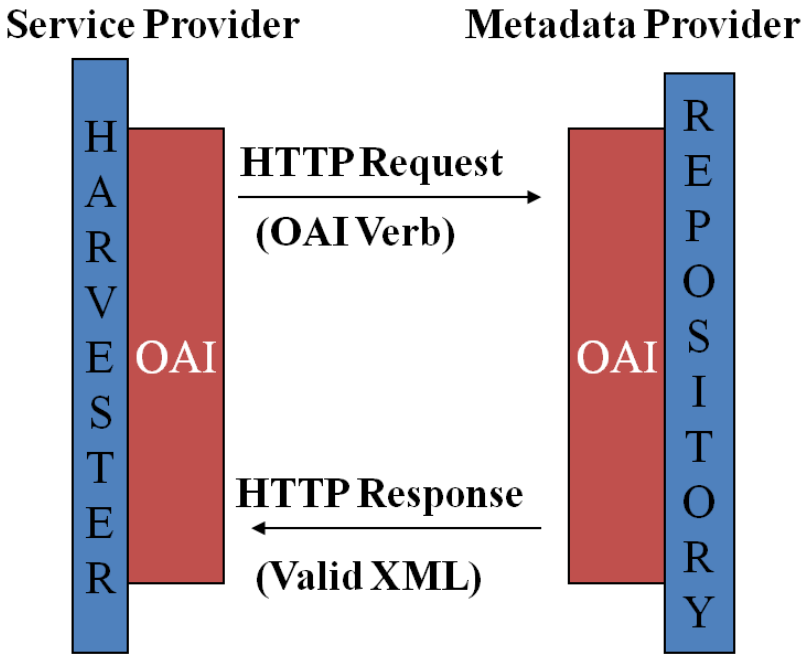
```
<record>
        <header>
                <identifier/>
                (format oai:repository id:record
                identifier)
                <setSpec/>
                <datestamp/>
        </header>
        <metadata>
                <prefix:format/>
                (Ex. oai_dc:dc) metadata in
                actual format
                </prefix:format>
        </metadata>
        <about> optional
                <provenance/>
                <rights/>
        </about>
</record>
```

Example OAI-PMH XML envelope

How OAI Works

OAI "VERBS"

- Identify
- ListMetadataFormats
- ListSets
- ListIdentifiers
- ListRecords
- GetRecord



europeana
carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

verb=Identify

Purpose

- Return general information about the archive and its policies (e.g., datestamp granularity)

Parameters

- None

Sample URL

- http://93.94.14.7/phpoai2/oai2.php?verb=Identify

europeana
carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

verb=ListSets

Purpose

- Provide a listing of sets in which records may be organized (may be hierarchical, overlapping, or flat)

Parameters

- None

Sample URL

- http://93.94.14.7/phpoai2/oai2.php?verb=ListSets

europeana
carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

verb=ListMetadataFormats

Purpose

- List metadata formats supported by the archive as well as their schema locations and namespaces

Parameters

- identifier – for a specific record (O)

Sample URL

- http://93.94.14.7/phpoai2/oai2.php?verb=ListMetadataFormats

europeana
carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

verb=ListIdentifiers

Purpose

- List headers for all items corresponding to the specified parameters

Parameters

- from – start date (O)
- until – end date (O)
- set – set to harvest from (O)
- metadataPrefix – metadata format to list identifiers for (R)
- resumptionToken – flow control mechanism (X)

Sample URL

- http://93.94.14.7/phpoai2/oai2.php?verb=ListIdentifiers&metadataPrefix=abm&set=KL

europeana
carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

verb=GetRecord

Purpose

- Returns the metadata for a single item in the form of an OAI record

Parameters

- identifier – unique id for item (R)
- metadataPrefix – metadata format for the record (R)

Sample URL

- http://93.94.14.7/phpoai2/oai2.php?verb=GetRecord&metadataPrefix=abm&identifier=oai:sffarkiv.no:SFFkl-100013

europeana
carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

verb=ListRecords

Purpose

- Retrieves metadata records for multiple items

Parameters

- from – start date (O)
- until – end date (O)
- set – set to harvest from (O)
- resumptionToken – flow control mechanism (X)
- metadataPrefix – metadata format (R)

Sample URL

- http://93.94.14.7/phpoai2/oai2.php?verb=ListRecords&metadataPrefix=abm&set=KL

europeana

carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

Unique Identifiers

Each item must have a unique identifier

Identifiers must follow rules for valid URIs

Example:

- oai:<archiveId>:<recordId>
- oai:www.carare.eu/1234567890
- (HTTP//):www.carare.eu/1234567890

Each identifier must resolve to a single item and always to the same item

- Can't reuse OAI item identifiers

europeana
carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

Datestamps

Needed for every OAI record to support incremental harvesting

Must be updated when addition or modification or deletion made in order to ensure changes are correctly propagated to harvesters

Different from dates within the metadata – OAI datestamp is used only for harvesting

Can be either YYYY-MM-DD or YYYY-MM-DDThh:mm:ssZ (must be GMT timezone)

CARARE is funded by the **European Commission's ICT Policy Support Programme**

Resumption Tokens

Use of resumptionTokens way to improve performance

- When transferring a big XML, instead of making it one file of X hundred MB, split it into chunks of N records
- Next time you issue a listRecords request and pass the resumption token, the transfer will start at N + 1 records
- Beneficial at the source – but also the destination (SAX or DOM parsing)
- resumptionTokens should retain state information for best performance

europeana
carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

Sets – option for selective harvesting

No well-defined semantics – depends completely on local data providers

- Must provide setSpec & setName, may provide setDescription, for each Set in repository

Sets may be hierarchical (use ":"); may overlap

- Allows for harvesting of sub-collections

May be pre-defined by arrangement between data providers and service providers

- E.g. Subject areas, years, author names (but must be pre-defined – for ListSets)
- Not a substitute for searching!

europeana

carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

Handling Metadata Record Deletions*

deletedRecord: no, transient, or persistent archives may keep track of deleted records, by identifier and datestamp

All protocol result sets can indicate deleted records (possible to delete a record, but not item)

Best Practice: If deletions are being tracked, this information should be stored indefinitely so as to correctly propagate to service providers with varying harvesting schedules

europeana

carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

Validation & Testing Tools

Services

- Repository Explorer (http://re.cs.uct.ac.za/)
- OAI Registry (http://www.oclc.org/oaister/)
- XML Schema Validator (http://www.w3.org/2001/03/webdata/xsv)

Command line harvesters:

- OAIHarvester2 (Java) download from http://www.oclc.org/research/activities/past/orprojects//harvester2/harvester2.htm
- UMharvester (Perl)

europeana

carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

Common Problems

Incomplete / inconsistent metadata

No unique identifiers

No datestamps

XML response not validating

europeana

carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

XML not validating

- Check namespaces and schema
- Use Repository Explorer in non-validating mode to check structure of XML, without looking at namespaces or schema
- Validate schema by itself if it is non-standard
- Look at XML produced by other repositories
- Watch out for character encoding issues

europeana
carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**