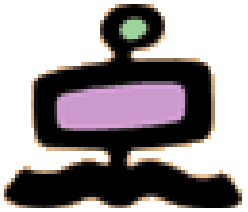
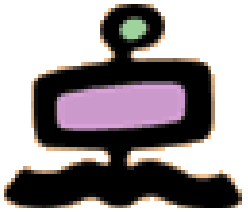
Corso di Biblioteche Digitali



- Vittore Casarosa
  - casarosa@isti.cnr.it
  - Office: 050 621 3115
  - Mobile: 348 397 2168
  - Skype: vittore1201

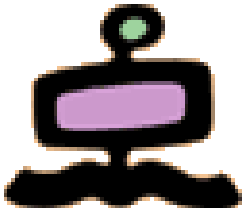
- “Ricevimento” at the end of the lessons or by appointment
- Final assessment
  - 70% oral examination
  - 30% project (development of a small digital library))
- Reference material:
  - Ian Witten, David Bainbridge, David Nichols, How to build a Digital Library, Morgan Kaufmann, 2010, ISBN 978-0-12-374857-7 (Second edition)
  - Material provided by the teacher
- **http://cloudone.isti.cnr.it/casarosa/BDG/**

Modules



- Computer Fundamentals and Networking
- A conceptual model for Digital Libraries
- Bibliographic records and metadata
- Information Retrieval and Search Engines
- Knowledge representation
- Digital Libraries and the Web
- Hands-on laboratory: the Greenstone system

Parallel evolution



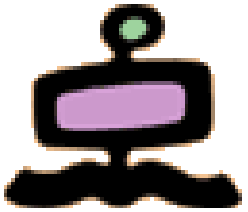
Libraries

- Description (documents)
  - Bibliographic records
  - MARC
- Interoperability
  - Z39-50
- Conceptual model (classes)
  - FRBR – LRM for Works, Expr., Manif.
- Information Retrieval
  - Full text (catalogue and documents)

The Web

- Description (instances)
  - Metadata
  - Dublin Core
- **Interoperability**
  - **OAI-PMH**
- Conceptual model (classes)
  - RDF and RDF Schema for all resources (ontologies)
- Information Retrieval
  - Full text (web pages and resources)

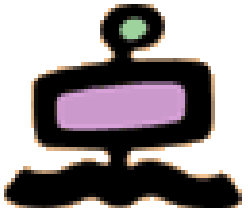
Exchange of information



Interoperability and the exchange of information

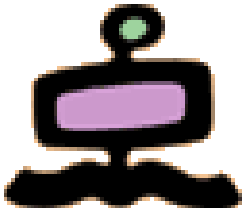
- The Z39.50 protocol
  - protocol defined to support the exchange of bibliographic data (basically MARC records) between libraries
- SRU/SRW queries
  - language defined to support “Z39.50 queries” carried over HTTP
- The OAI-PMH protocol
  - protocol defined to support the exchange of metadata (Dublin Core and others)

OAI – Open Archives Initiative



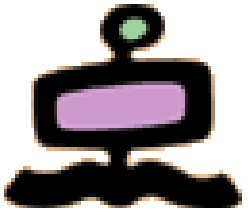
- The roots of OAI lie in the development of eprint archives (i.e. Institutional Repositories) such as arXiv, CogPrints, NACA (NASA), RePEc, NDLTD, NCSTRL, etc.
- Each repository offered a web interface for deposit of articles and for end-user searches
- Users working across multiple archives had to learn multiple different interfaces
- Need to define a single interface to search across multiple repositories
- Initial experiments for single search interface to a number of “federated archives” (UPS - Universal Pre-print Service)
- UPS renamed OAI (Open Archives Initiative) at the Santa Fe Convention (1999)

Institutional Repositories



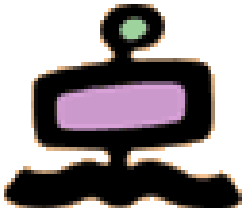
- An Institutional repository is a centrally managed collection of institutionally generated digital objects designed to be maintained “for ever”
- Established and maintained by universities and research institutions (initially) to cope with self-publishing
- An *e-print* is an author self-archived document. The content of an e-print is usually the result of scientific or other scholarly research.
- Repositories contain scholarly publications
  - Reports
  - Working papers
  - Pre- and post-prints of articles and books
  - Doctoral thesis
  - Data supporting research
  - References and professional databases related to research topics

OAI – Open Archives Initiative



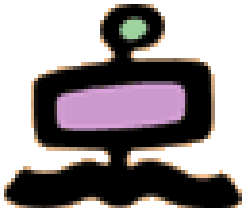
- The roots of OAI lie in the development of eprint archives (i.e. Institutional Repositories) such as arXiv, CogPrints, NACA (NASA), RePEc, NDLTD, NCSTRL, etc.
- Each repository offered a web interface for deposit of articles and for end-user searches
- Users working across multiple archives had to learn multiple different interfaces
- Need to define a single interface to search across multiple repositories
- Initial experiments for single search interface to a number of “federated archives” (UPS - Universal Pre-print Service)
- UPS renamed OAI (Open Archives Initiative) at the Santa Fe Convention (1999)

Z39.50 and OAI



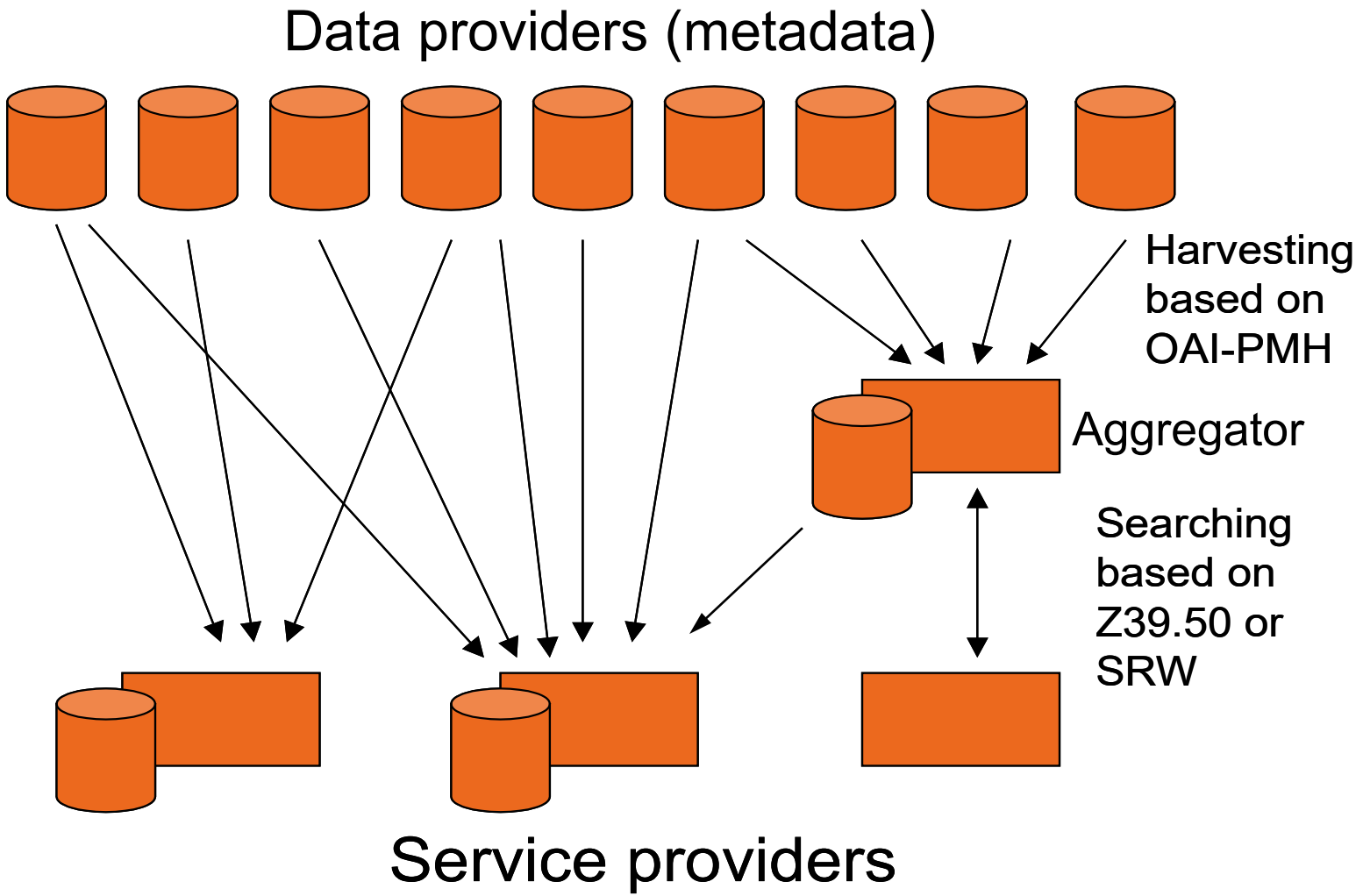
- It is interesting to compare OAI and Z39.50, two protocols for “resource discovery”
- Historical separation from Z39.50
  - OAI appears about 15 years after Z39.50
- Cultural separation from Z39.50
  - Z39.50 originated in the traditional library community
  - OAI originated in the “Web Community”
- Conceptual separation from Z39.50
  - Z39.50 based on solid (but heavy and bulky) foundations
  - OAI based on simple and pragmatic ideas

Searching versus Harvesting

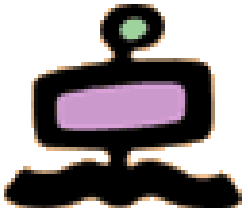


- Two possible approaches for single search interface to multiplel archives
  - cross searching multiple archives based on “query replication” (like Z39.50, possibly lighter)
  - harvesting metadata into one or more ‘central’ services
- Problems with cross searching (query replication)
  - Not scalable (overall performance determined by slowest server)
  - Problems of deciding which servers to target (collection descritpions not consistent)
  - Differences in interfaces and query languages
  - Problems in the merging of ranked results (different types and size of targets can skew results)
  - Browse interface very difficult to build
- Decision was to go with harvesting

OAI – PMH overview



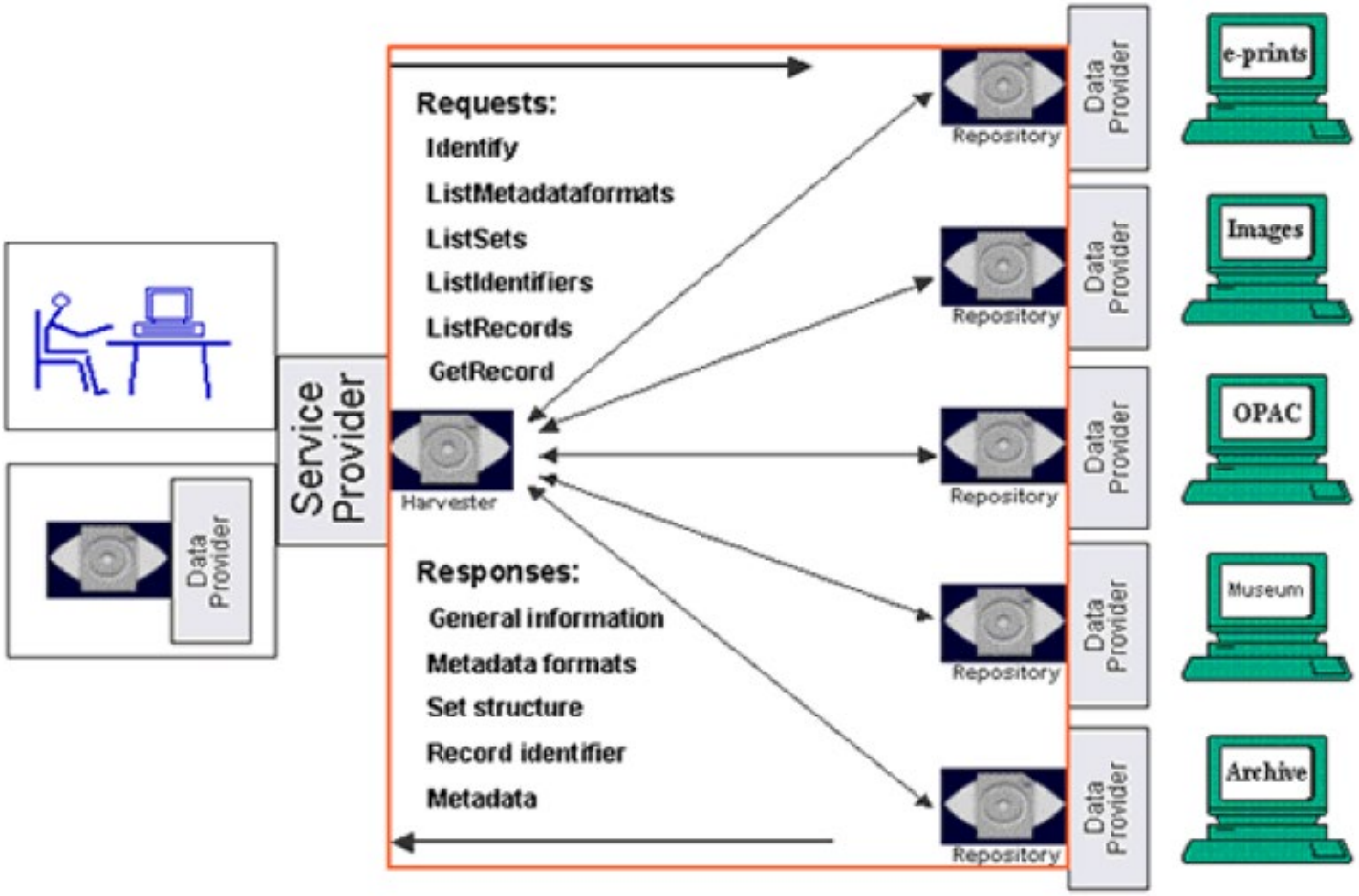
OAI - PMH



OAI Protocol for Metadata Harvesting

- Data Providers make metadata available for harvesting
- Service Providers harvest metadata
- Data Providers
  - Are creators and keepers of metadata (repositories) and (possibly but not necessarily) archives of resources (the objects described by the metadata)
  - Handle deposit and publishing of metadata and (possibly) digital objects
- Service Providers
  - Are harvesters of metadata for the purpose of providing a service such as a search interface, peer-review system, etc.

Overview of OAI - PMH



Conceptual model of OAI data

resource

item = identifier

all available metadata about *David*

item

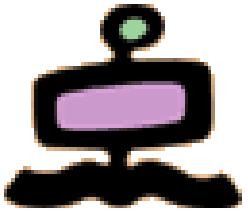
Dublin Core metadata

MARC metadata

SPECTRUM metadata

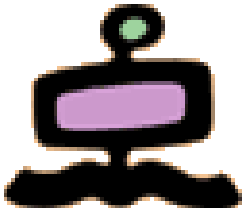
records

OAI-PMH Protocol Overview



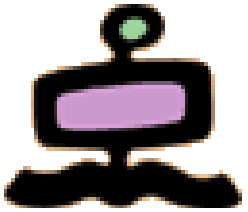
- Protocol based on HTTP
- Request arguments as GET or POST parameters
- Six request types (verbs)
- Responses are encoded in XML syntax
- Supports any metadata format (Dublin Core mandatory for each data provider)
- Support selective harvesting
  - logical set hierarchy (data providers)
  - date stamps (last change of metadata set)
- Flow control (token to retrieve subsequent records)
- Error messages

OAI – PMH verbs



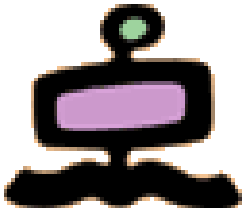
- Identify
  - description of an archive
- ListMetadataFormats
  - retrieve available metadata formats from archive
- ListSets
  - retrieve set structure of a repository
- ListIdentifiers
  - abbreviated form of ListRecords, retrieving only headers
- ListRecords
  - harvest records from a repository
- GetRecord
  - retrieve individual metadata record from a repository

Logical set hierarchy (sets)

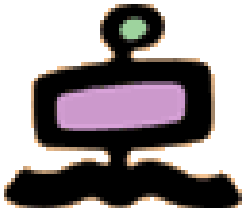


- Logical partitioning of repositories
- Optional – archives do not have to define sets
- Useful to support selective harvesting
- Sets can be defined according to different criteria
  - publication date
  - publication types (thesis, article, …)
  - document types (text, audio, image, …)
  - subject, (possibly) according to controlled vocabularies (possibly) in different fields (medicine, biology, computer science, …)

Flow control

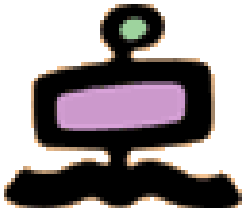


OAI – PMH records



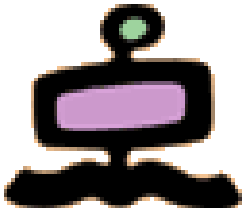
- A record contains the metadata of a resource in a specific format
- It has three parts
  - header (mandatory)
    - identifier
    - datestamp
  - metadata (mandatory)
    - XML encoded metadata with root tag, namespace
    - repositories MUST support Dublin Core
    - MAY support other formats
  - about (optional)
    - rights statements
    - provenance statements

OAI – PMH request



- Requests must be submitted using the GET or POST methods of HTTP
- Repositories must support both methods
- At least one key=value pair: verb=[RequestType]
- Additional key=value pairs depend on request type
- Example for GET request
  - http://archive.org/oai?
    verb=ListRecords&
    metadataPrefix=oai_dc

OAI – PMH example



http://edoc.hu-berlin.de/OAI-2.0?

verb=ListIdentifiers&

from=2002-01-06&

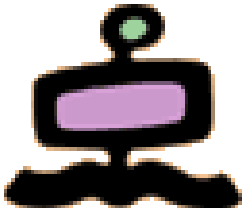
until=2002-01-08&

metadataPrefix=oai_dc&

set=doctypes:dissertations

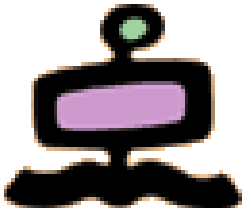
- ListIdentifiers returns the record headers

OAI – PMH response



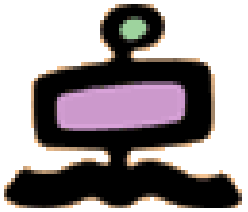
- Formatted as HTTP responses
- Content type must be text/xml
- HTTP lossless data compression optional
- XML declaration
  (<?xml version="1.0" encoding="UTF-8" ?>)
- Root element named OAI-PMH with three attributes (xmlns, xmlns:xsi, xsi:schemaLocation)
- Three child elements
  - ResponseDate (UTC datetime)
  - Request (copy of the request that generated the response)
  - a) error (in case of an error or exception condition)
  - b) element with the name of the OAI-PMH request

Response to ListIdentifiers (1/2)



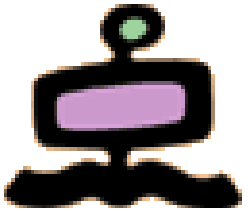
```
<?xml version="1.0" encoding="UTF-8"?>
    <OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
                xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
                xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
                            http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
       <responseDate>2002-10-22T17:49:49+01:00</responseDate>
       <request verb="ListIdentifiers" from="2002-01-03" until="2002-01-08"
                            metadataPrefix="oai_dc" set="doctypes:dissertations">
                http://edoc.hu-berlin.de/OAI-2.0</request>
       <ListIdentifiers>

        ......   details in a moment

       </ListIdentifiers>
    </OAI-PMH>
```

Response to ListIdentifiers (2/2)



```
<ListIdentifiers>
        <header>
            <identifier>oai:HUBerlin.de:3000819</identifier>
            <datestamp>2002-01-08</datestamp>
            <setSpec>doctypes</setSpec>
            <setSpec>doctypes:dissertations</setSpec>
            <setSpec>dnb</setSpec>
            <setSpec>dnb:dnb33</setSpec>
        </header>
        <header>
            <identifier>oai:HUBerlin.de:3000831</identifier>
            <datestamp>2002-01-07</datestamp>
            <setSpec>doctypes</setSpec>
            <setSpec>doctypes:dissertations</setSpec>
            <setSpec>dnb</setSpec>
            <setSpec>dnb:dnb27</setSpec>
        </header>
    </ListIdentifiers>
```

What we have seen



- Digital Libraries
  - Discovery of information
    - Describing Information
      - Metadata
        - MARC
        - Dublin Core
        - MODS
        - METS
        - TEI
        - EAD
        - .....
      - Knowledge Representation
        - FRBR
        - RDF
    - Interoperability
      - Queries
        - Z39.50 queries
        - Common Command Language (CCL – ISO 8777 or Z39.58)
      - Protocols
        - Z39.50
        - SRU/SRW
        - OAI-PMH