Assessing Open Archive OAI-PMH implementations

Emanuele Bellini, Marcel Aime Deussom, Paolo Nesi

Distributed Systems and Internet Technology Laboratory
University of Florence, Department of Systems and Informatics
http://www.disit.dsi.unifi.it nesi@dsi.unifi.it

Rationales

- Diffusion of Open Access in terms of service adoption and political pushing, vehicle for diffusing research results
  - Initiatives supporting and setting up Open Archives are growing and have adopted OAI formats and solutions
  - Actions are going to be taken to increase diffusion, etc.
- OAI-PMH
  - protocol to give access to archive records
  - servers provide collected data related to the records of their collections
- Presently there are about 2000 listed OAI-PMH compliant institutional repositories/servers in the world
  - Probably many others are presented and not listed.
  - Millions of records are stored in this distributed architecture

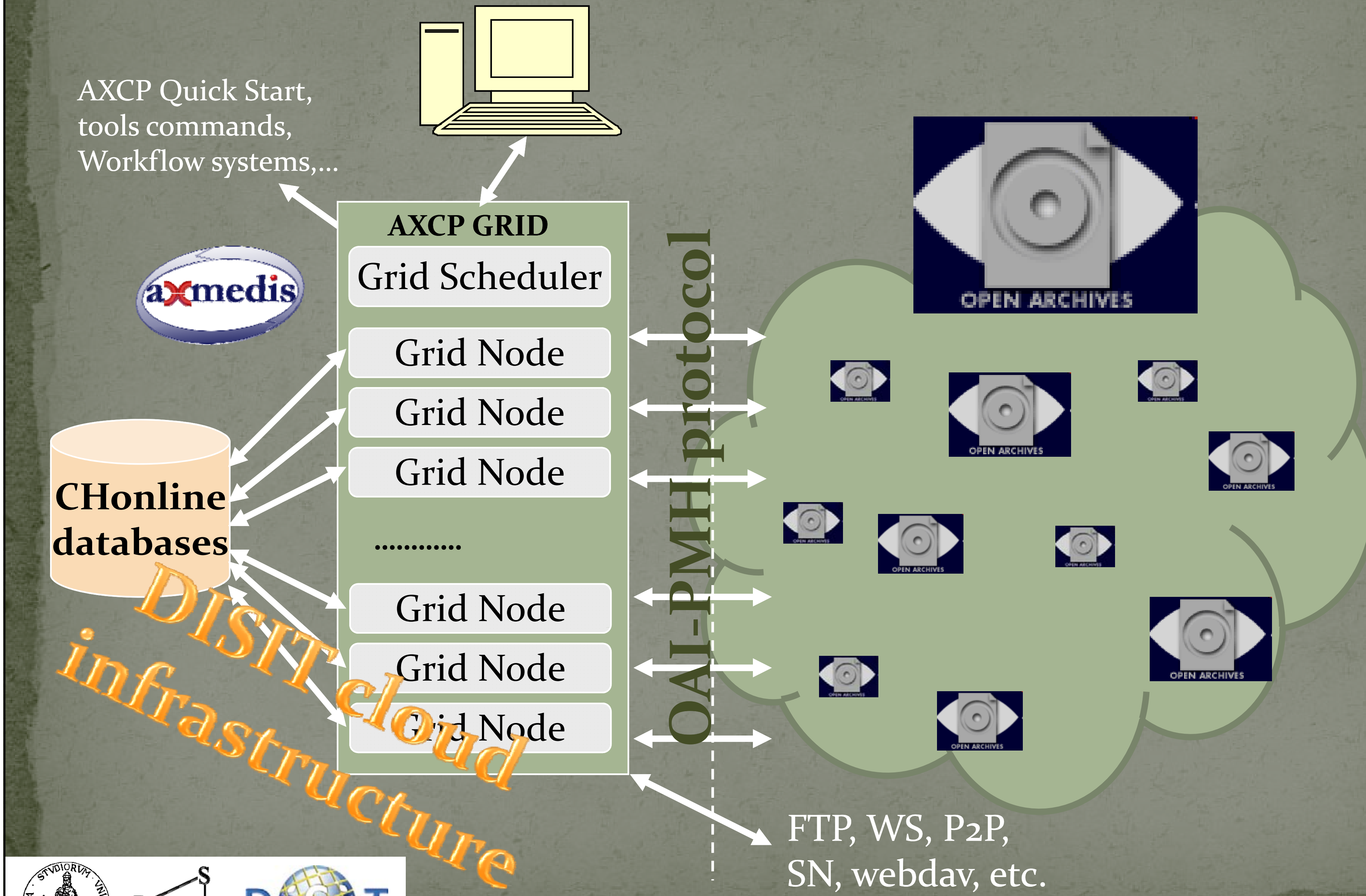
Objectives

- The amount of OAI-PMH repositories and their interoperability is at the basis of the architecture
  - Servers provide access to their records in different formats
- Suddenly a lot of problems have been neglected or under estimated in this architecture
  - They prevent the integration of data coming from different archives
- The most relevant are the
    - Quality of the *OAI-PMH repository service*
    - Quality of the Metadata provided
  - Quality, in this case, mainly has to take care about:
    - Completeness, Accuracy, Consistency, compliance, ..Reliability, ..
    - ...see ISO standards ...

Early assessment of complexity

- Started from www.openarchive.org list
- Realized that
  - A large number of repositories are
    - not working in the correct manner, as explained in the following
    - providing Archival records in multiple Metadata models
  - Metadata models are
    - not 100% compliant with their definition
  - OAI-PMH protocol
    - does not include any way to know the total number of records in advance
    - is quite slow for massive harvesting
    - records related to the same content in different languages are typically not related each other
  - The total number of records is larger than 18 millions, including duplications of different kind… → actually >36 Millions

Massive Harvesting Architecture



AXCP Quick Start, tools commands, Workflow systems,…

AXCP GRID

Grid Scheduler

Grid Node

Grid Node

Grid Node

…………

Grid Node

Grid Node

Grid Node

CHonline databases

DISIT cloud infrastructure

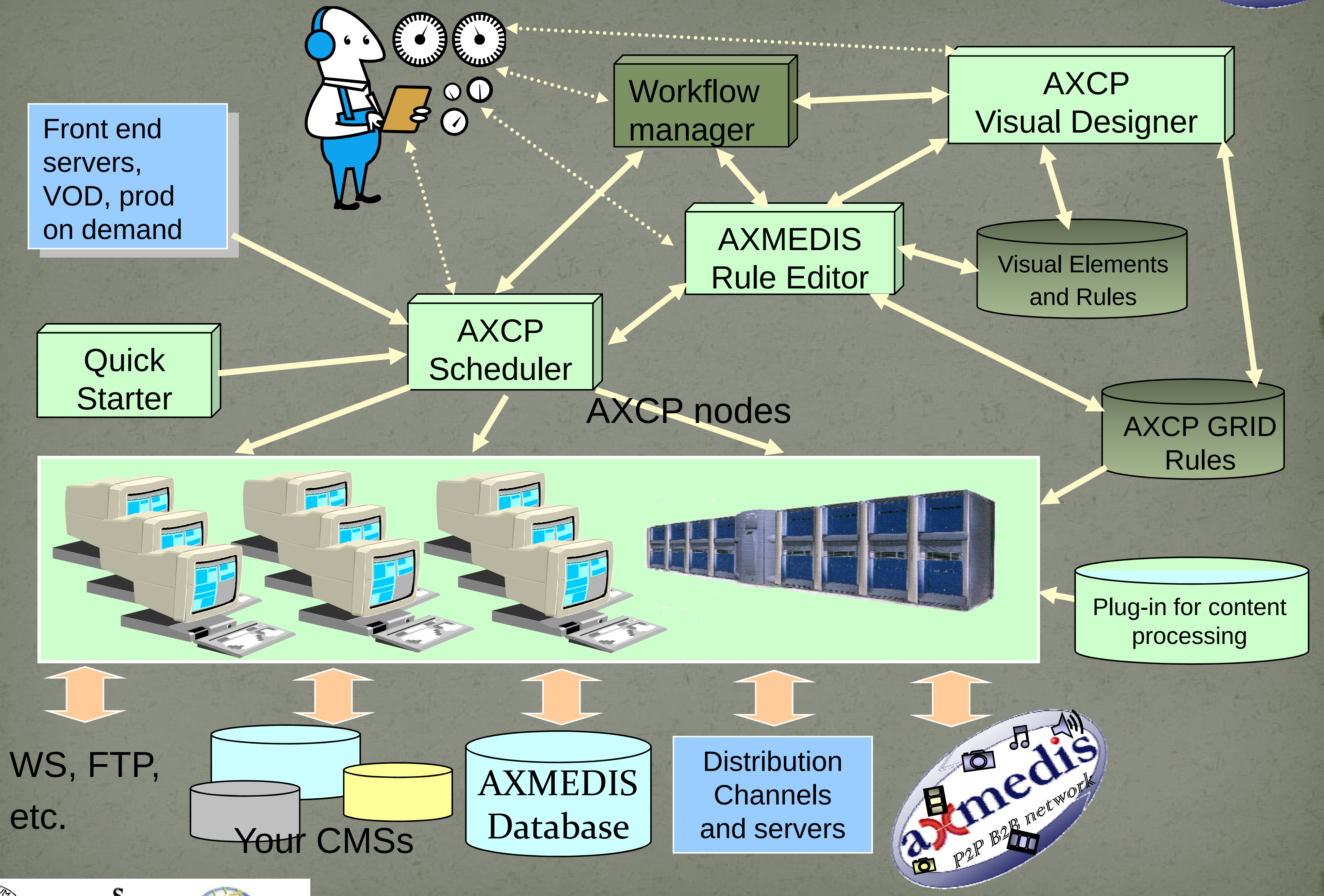
OAI-PMH protocol

FTP, WS, P2P, SN, webdav, etc.

Media GRID Implementation details

- AXMEDIS AXCP grid tool has been used
  - developed at DISIT on AXMEDIS IP EC R&D project
  - freely distributed on hundreds of internet portals
- GRID Node Processes are
  - defined/programmed in Extended JavaScript by using an IDE for development with debug and execution capabilities
  - automatically allocated on grid nodes by Scheduler
- Scheduler and Nodes can be allocated on Virtual machines on a Cloud infrastructure: 15-30 nodes
- *Access to OAs via OAI-PMH is performed via internet and has to be capable to recover faults of several kind:*
  - *broken connections, stopped server, etc.*

AXMEDIS Content Processing media GRID



Front end servers, VOD, prod on demand

Workflow manager

AXCP Visual Designer

AXMEDIS Rule Editor

Visual Elements and Rules

Quick Starter

AXCP Scheduler

AXCP nodes

AXCP GRID Rules

Plug-in for content processing

WS, FTP, etc.

Your CMSs

AXMEDIS Database

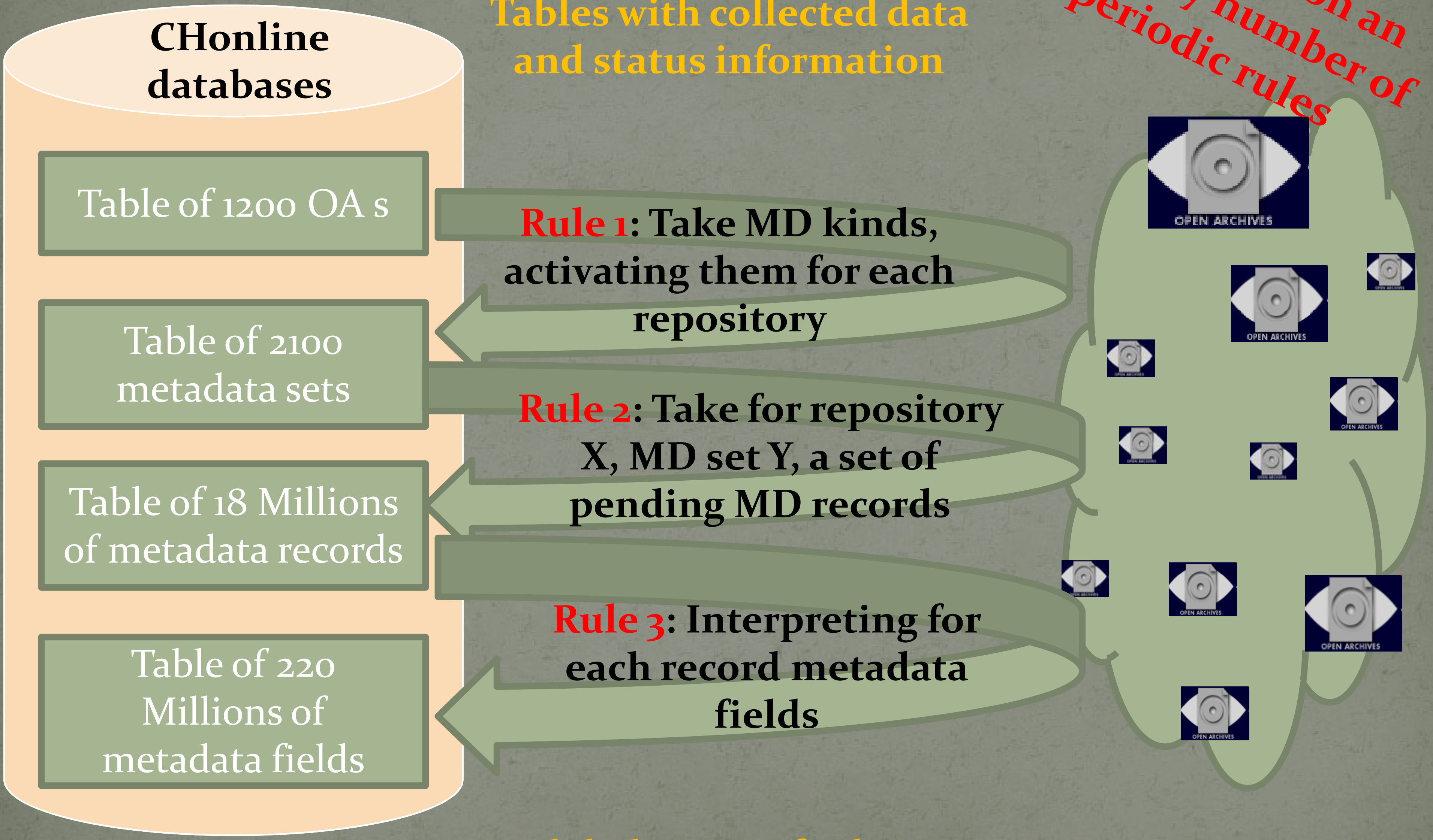
Distribution Channels and servers

AXCP Scalable Back Office

- Automating and scaling up:
  - Content ingestion and integration, database management, etc.;
  - Content processing, formatting, adaptation, transcoding, etc.;
  - Metadata mapping and processing; by direct mapping and semantic reasoning
  - harvesting and crawling via OAI-PMH, P2P, HTTP, etc.
    - Connection with other social networks such as YouTube and Flick to propagate queries and get content or for posting content;
  - Content and users similarity analysis and clustering, for users and content recommendations; this processing has to be performed off line due to its computational complexity. Results will be immediately usable by the users to identify similar content and users as described above.
  - Content aggregation and integration (packing, packaging) for educational and entertainment productions;
  - multilingual processing, text processing, semantic processing,
  - conversion from XML to RDF
  - ..

Processing workflow

Delegating on an arbitrary number of periodic rules



CHonline databases

Table of 1200 OA s

Table of 2100 metadata sets

Table of 18 Millions of metadata records

Table of 220 Millions of metadata fields

Tables with collected data and status information

Rule 1: Take MD kinds, activating them for each repository

Rule 2: Take for repository X, MD set Y, a set of pending MD records

Rule 3: Interpreting for each record metadata fields

Modeled in a unified manner

Assessment model and analysis

- To understanding problems and weakness at both quantitative and qualitative levels.
  - **General Metrics (GenMet)**
    - quantitative evaluation outlining a global view of the diffusion and effective available OAI-PMH implementation services;
  - **Archive Metrics**
    - quality evaluation related to the usage and trustability of adoption of metadata sets;
  - **Metadata Record Metrics**
    - modeled and provided according to the adopted standard, assessing the usage of single metadata fields and their potential weakness.

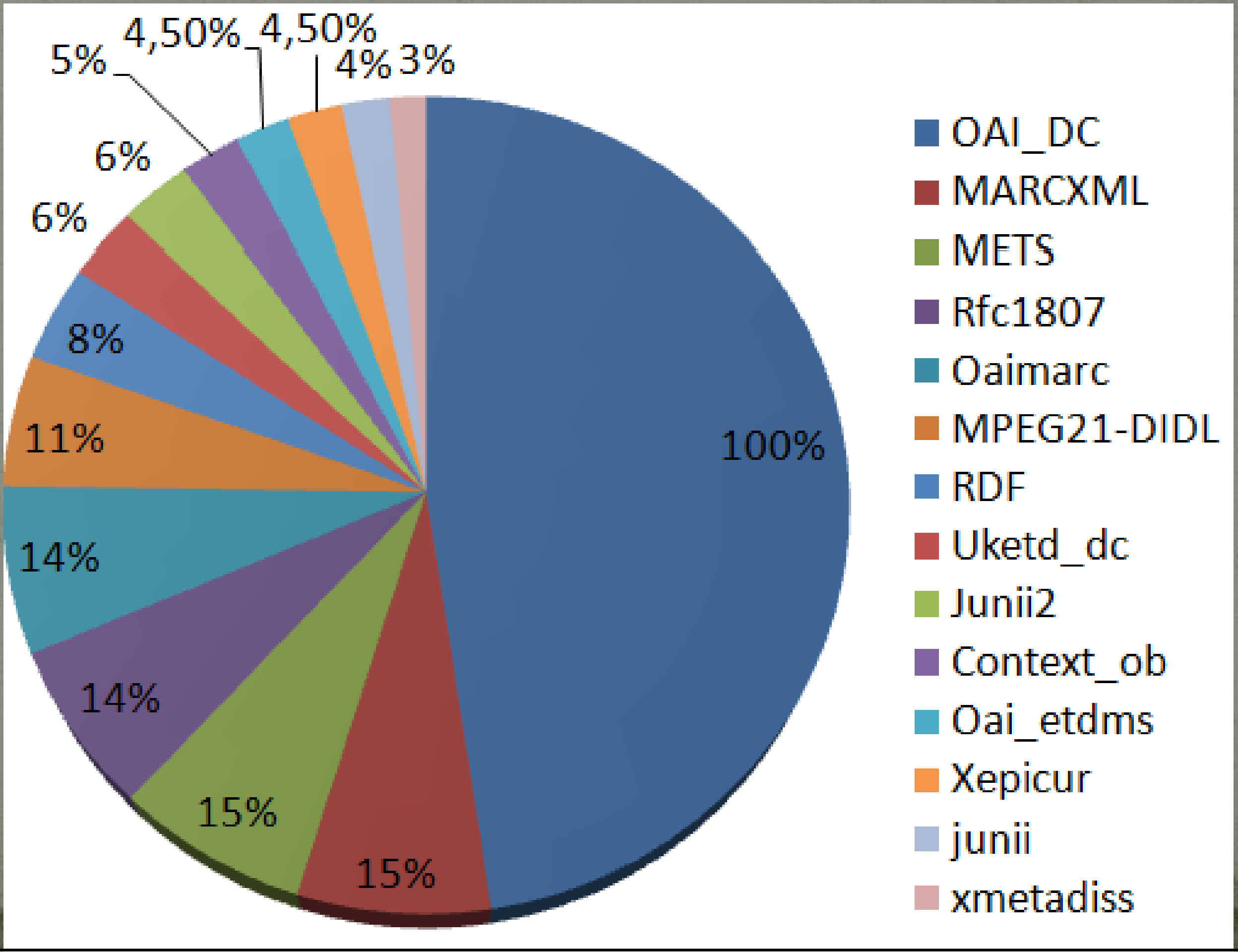
GenMet: OA available for harvesting

- Sometime Official OAs are not accessible
  - On the 1200 identified 1200 OAs,
    only 74% of them where active providing records

- Open Archives where Not accessible since:
  - They have been just set up as experiments with a few number of records, and put off line after that
  - High costs for maintaining active and OA,
    - underestimation of those costs.
    - lack of direct connection with the central archive

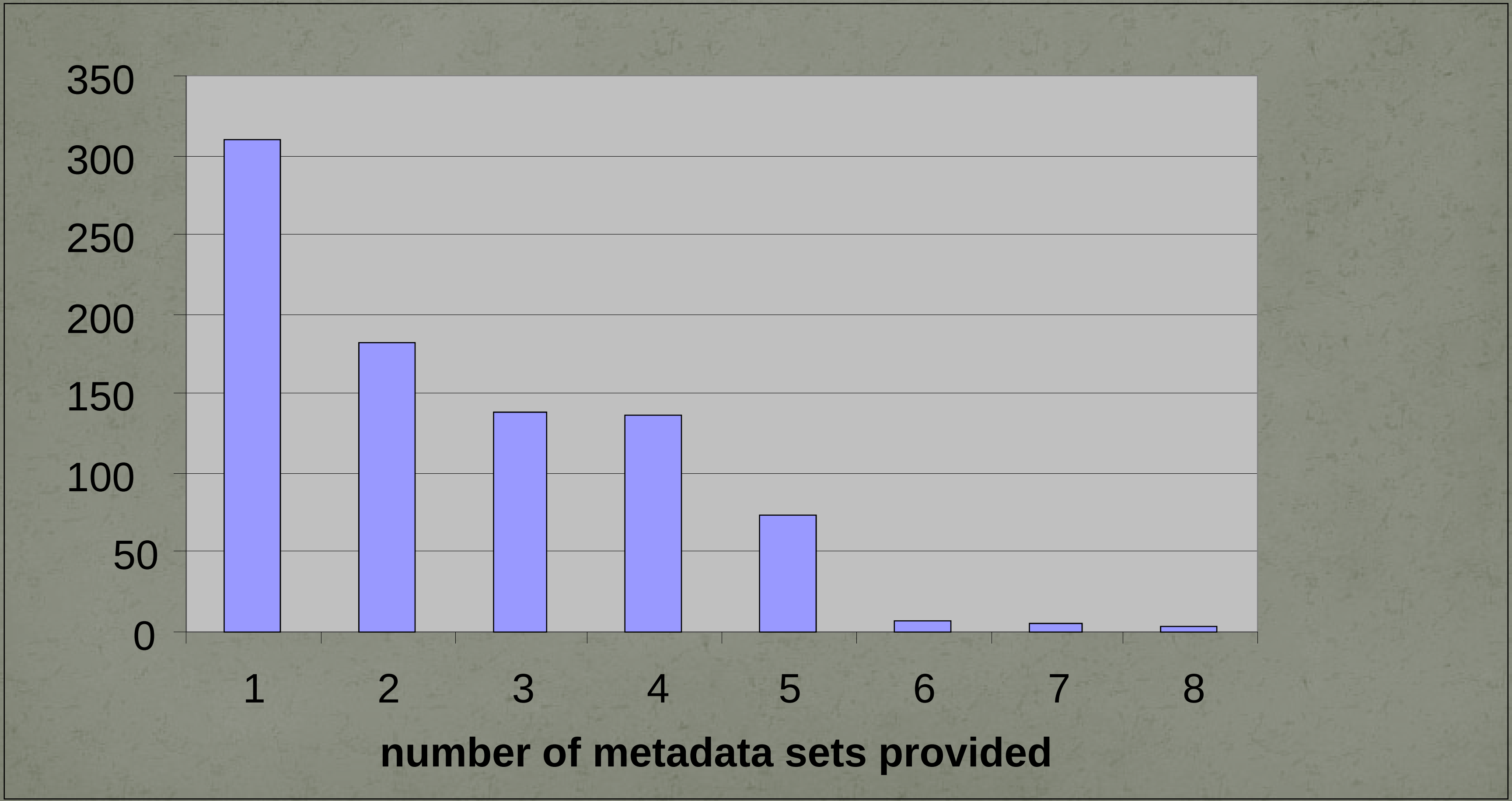
GenMet: OA Metadata sets

- According to the OAI-PMH protocol, each OA declares to have a number of metadata sets available.
  - The 10% of them have inconsistency between the declared and provided metadata sets.
- A total of 153



Metadata sets managed by single OA

- Distribution of Metadata sets provided by OAs



Archive Level Metrics

- Some of them, do not return correct information
  - E.g., returning HTML instead of XML !
- Reference Schema for the Metadata set
  - The references to schemas are frequently wrong or not the official links
- Use of Metadata set in the Archive
  - 15% use metadata sets which are personalized and/or custom made
- Small Open Access Archives
  - 14% have less than 100 records
- Empty Metadata Records in the Archive
  - 84% have no empty records
  - 9% have less than the 5% of empty records

Metadata Record Metrics

- Mainly due to wrong interpretation of MD model
- For Dublin Core, for example:
  - Date: wrong date or format, hard to interpret
    - E.g.: Agust 2011, settembre 1987, middleage, 01/03/02
  - Author: wrong or bad formatted
    - E.g.: A. Rossi, Frank J. Cross, Paolo Nesi; P. Nesi, Nesi,
  - Rights: a free text vs formal REL (weak definition)
  - Format and/or Type: range from MIME type to simple file extension or textual description
    - E.g.: Wrong extension, wrong mimetype IANA, Text description
    - In some cases also the Filename
    - Wrong in the 10% of cases.
  - Identification:
    - From file name to some GUID, URI is any, etc., broken links, etc.

DC: language coding

- Several different versions and coding, some wrong coding / writing

language	instances	#
English	en, eng, English, en_GB, en-GB, Englisch	6
Spanisch	es, spa, Espanol, Spanish, spa, sp	6
French	fr,fre, French, French, Francais, fra	6
Deutsch	ger ,de, German, Deutsch, ge	5
Greek	gr, gre, grc, ell	4
Italian	it,ita, Italian	3
Japan	jpn, ja, jp	3

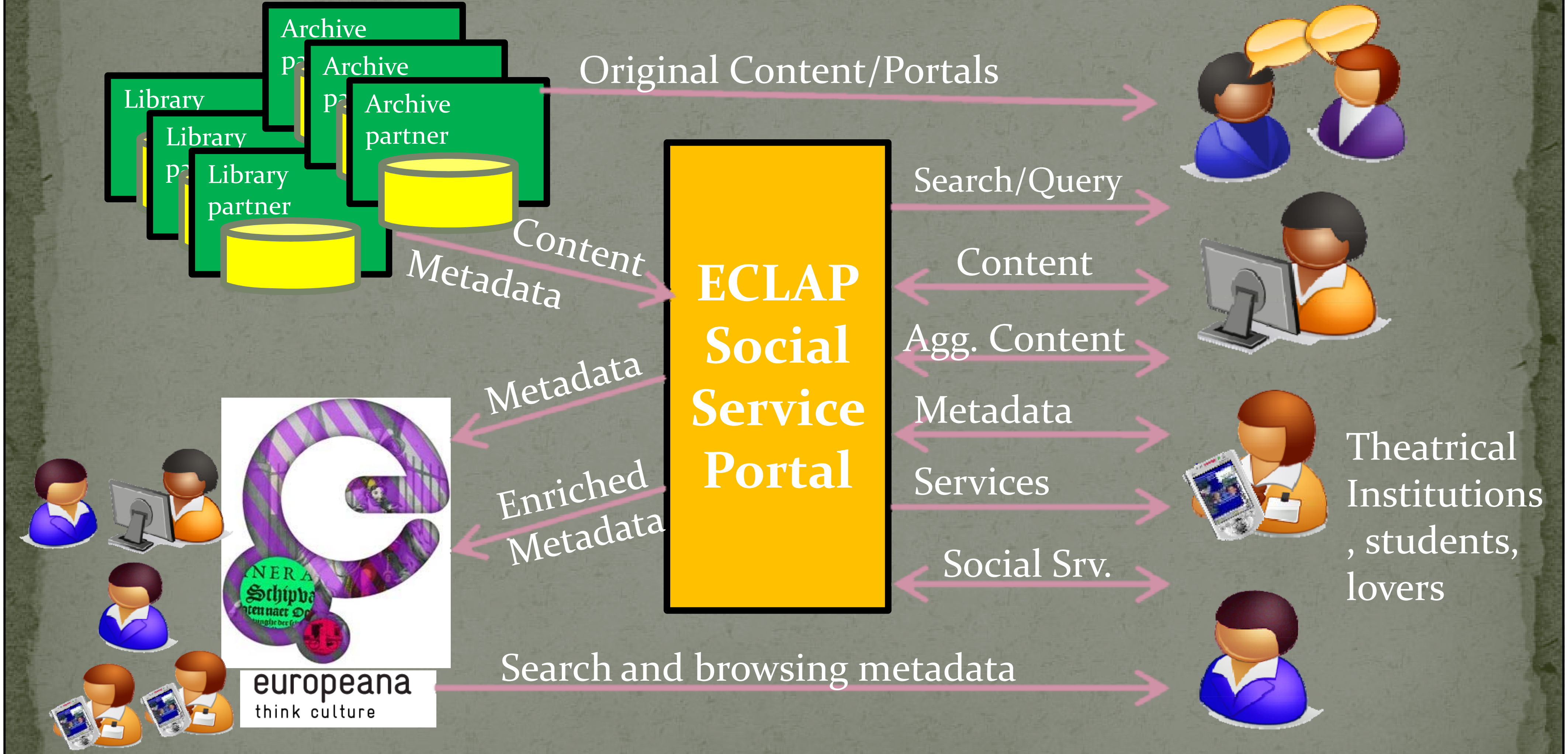
Quality of Metadata Fields

- Poor quality creates
  - problems of interoperability
  - Problems in time ordering, querying…
  - difficulties in detecting duplications
  - difficulties in detecting cases in which the MD information can be merged to obtain more complete sets.
- Solutions:
  - some error recovering/correction is possible, to integrate data in the same unified archive, and enabling queries
- This is What we are going to do in **ECLAP project**.

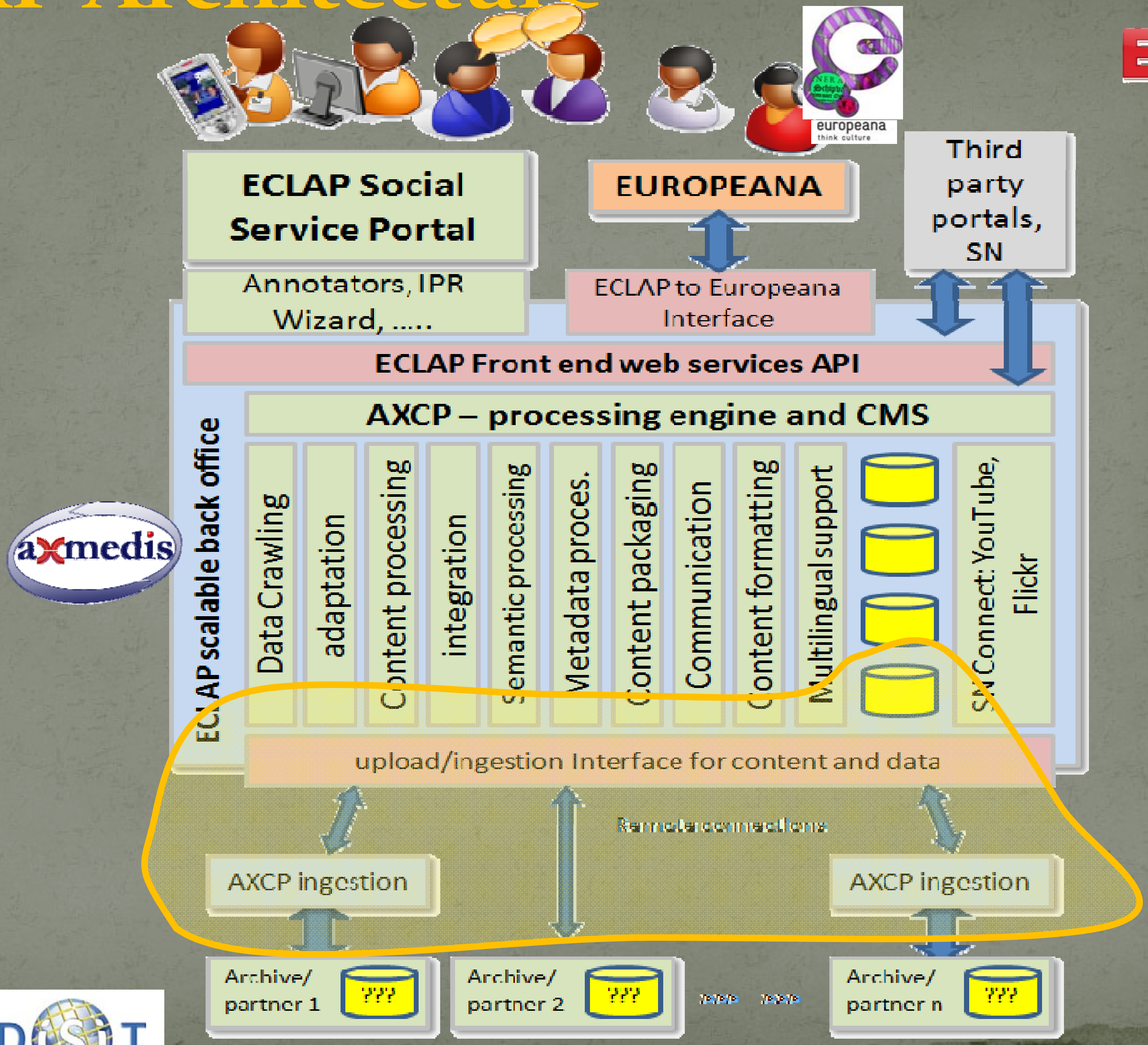
ECLAP CIP PSP EC Project, aim

http://www.eclap.eu

http://bpnet.eclap.eu *Best Practice Portal*



ECLAP Architecture



ECLAP Social Service Portal

Annotators, IPR Wizard, …..

EUROPEANA

ECLAP to Europeana Interface

Third party portals, SN

ECLAP Front end web services API

ECLAP scalable back office

AXCP – processing engine and CMS

Data Crawling

adaptation

Content processing

integration

Semantic processing

Metadata proces.

Content packaging

Communication

Content formatting

Multilingual support

SN Connect: YouTube, Flickr

upload/ingestion Interface for content and data

Remote connections

AXCP ingestion

AXCP ingestion

Archive/ partner 1 ???

Archive/ partner 2 ???

Archive/ partner n ???

Partners and third party Archives in OAI, XML, ODBC, DC, EDM, ….

Conclusions

- Definition and application of an **assessment model**
  - Definition of a set of metrics at different OA levels
  - Application of metrics on 1200 OA repositories to assess the general quality level in the OA world
  - Identification of the most common problems in managing and producing metadata for OA solutions
  - Assessment of the world wide level of quality for OAs
- Realization of an ingestion AXCP Grid Based tool
  - To automatically ingest huge amount of data , and
    - integrate all MD sets in a common model
  - It can be also used to update, process, correct, repurpose..
- The **Assessment Tool for your OA** will be available for all in connection to *ECLAP CIP PSP* project of the European Commission.