Harvesting dei Metadati negli Archivi Italiani

Realizzazioni e problemi

Zeno Tajoli – CILEA. tajoli@cilea.it

Tecnologia usata

- Linguaggio di riferimento: Perl 5.8.5
- Harvester: Celestial 2.1.5 con aggiunti:
  - Scarico solo struttura archivio
  - Selezione Sets
- Crosswalks: modulo XML::Twig 3.1.15
- Indicizzazione: Cheshire 2.39g

Struttura generale

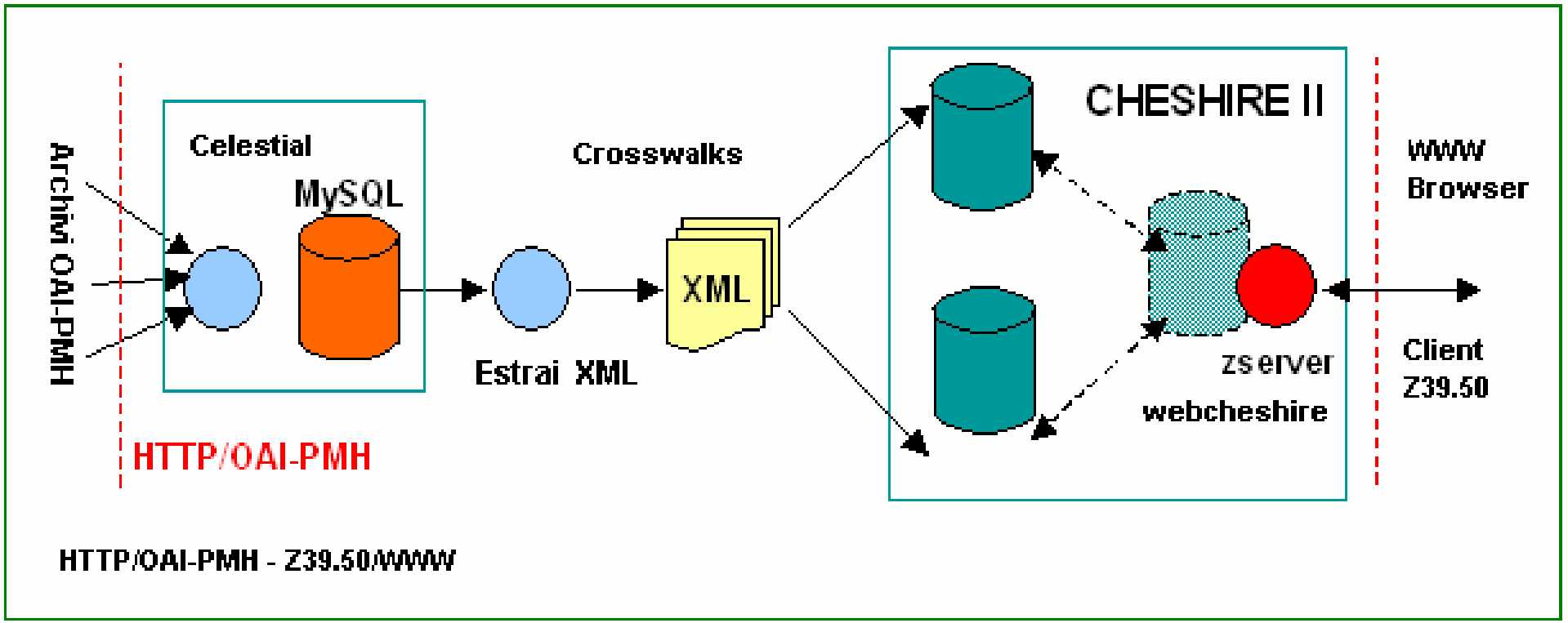

Immagine tratta da RDN

Criteri di selezione

- Supporto del protocollo OAI-PMH
- Ambito italiano
- Ambito istituzionale
- Metadati collegati a un full-text liberamente disponibile

Archivi selezionati e presenti

- Università di Bologna (Acta e Miscellanea)
- Università di Firenze
- Università di Padova
- Università di Trento
- Sissa (Trieste)
- E-LIS (solo italiano)
- CNR di Bologna

Archivi in lavorazione

Si sta affrontando la rete RePEc <http://repec.org>
Di essa fanno parte, ad esempio :

- Banca d'Italia
- Fondazione Eni Enrico Mattei
- Università di Firenze, Dipartimento di Statistica
- Istituto Innocenzo Gasparini (Bocconi)
- Università dell'Insubria (Facoltà di Economia)
- Libero Istituto Universitario Carlo Cattaneo
- Dipartimento di Economia della Statale di Milano

…

Archivi non utilizzati

- Contributi italiani non selezionabili (es. arXiv)
- Didattici (es. DSpace@unipr)
- Istituzionali non italiani (es. Istituto Universitario Europeo)

Full – Text disponibili (22/10/2004)

• Università di Bologna (Acta)	764
• Università di Bologna (Miscellanea)	22
• Università di Firenze	265
• Università di Padova	7
• Università di Trento	461
• Sissa (Trieste)	577
• E-LIS (solo italiano)	383
• CNR di Bologna	8

Il formato usato

Sono usati questi indicatori

- dccreator
- dctitle
- dcdescription
- dcidentifier
- dcsubject
- dctype
- dcformat
- dcrights
- dcarchive
- dclanguage
- dcrelation

Significato degli indicatori

Gli indicatori derivano da Dublin Core, aggiunti:

- “dcarchive”
- “dcrelation”

Interventi sui dati

- dccreator: qui confluiscono dc:creator e dc:contributor
- dclanguage: sigle in ISO 639-2
- dcdate: qui la data riferita alla pubblicazione del full-text

Interventi sui dati

- dctype: le tipologie accolte sono:
  - Articles/Journals
  - Books/Chapters
  - Conference Papers
  - Dissertations
  - Grey Papers
- dctype: registra delle sigle nei dati
- dc:publisher: viene cancellato

Interventi sui dati

dcsubject:

- classificazione con materie MIUR (usando sigle)
- Operazioni di conversione per chi non la supporta

Problemi rilevati:

- Non sempre presente il dato, usati valori fissi
- I soggetti possono essere più generali delle classi usate
- Sono possibili incongruenze

Problematiche dei dati

- La necessità di usare sigle in dcsubject e dctype
- L'indicatore dcrelation è raro
- Incongruenze in dcformat

Problematiche dei dati

- I sets sono implementati in maniera abbastanza difforme
- Mancano le keywords
- Come gestire la presenza di più archivi istituzionali per lo stesso ente ?

Suggerimenti

- Un formato comune [itarchiveformat ?]
- Una classificazione comune
- Un accordo per dc:type
- Un accordo sull'implementazione dei sets