

1st Plone for Research and University Day - Bologna, 20 Maggio 2010

# Glossari, thesauri, ontologie: annotazione e ricerca dei contenuti su base semantica

*Giovanni Toffoli - LINK srl, Roma*

## CLASSIFICAZIONE E RICERCA MEDIANTE KEYWORD - 1

Plone consente di classificare i contenuti associando una o più **keyword** a ciascuno di essi.

- nelle recenti versioni di Plone le keyword sono chiamate **categorie**.

Le keyword, o categorie, fanno parte dei **metadati** standard di Plone

- le keyword costituiscono il valore del campo **Subject**, che corrisponde all'omonimo campo nello standard **DublinCore**.

## CLASSIFICAZIONE E RICERCA MEDIANTE KEYWORD - 2

Il catalogo di Plone indicizza le keyword nell'indice **Subject**:

- consente di ricercare tutti i contenuti che siano stati classificati mediante almeno una delle keyword specificate in una query, o anche tutte.

Chi decide quali keyword si possono usare in un sito?

- si può dare la massima libertà a tutti
- oppure un "amministratore" può definire e mantenere un repertorio di keyword ammesse
- ma si può anche seguire una via di mezzo: lasciare una libertà di base, ma periodicamente sfoltire il repertorio; in questo è di aiuto PloneKeywordManager, un'estensione semplice ma utile e di facile uso:

<http://plone.org/products/plonekeywordmanager>

## CLASSIFICAZIONE E RICERCA MEDIANTE KEYWORD - 3

Le keyword in linea di principio sono distinte dal contenuto testuale di un documento, anche se ovviamente possono ispirarsi ad esso.

Le keyword sono affini ai **descrittori** o **termini controllati** largamente usati in **biblioteconomia**.

Descrivere un contenuto mediante keyword è come dargli una **collocazione** concettuale in aggiunta a quella fisica; con la differenza che tale collocazione può essere multipla.

## LA RICERCA SEMANTICA - 1

Il termine **semantico** si usa per lo più per riferirsi al contenuto "profondo" di un oggetto, di un documento; al suo "significato", contrapposto al suo aspetto di superficie.

In realtà una persona che ha pratica di ricerca **full-text**, anche se cerca per parole, di solito ottiene un'alta percentuale di risultati che ben realizzano il "concetto" che egli ha in testa.

Anche se il **linguaggio naturale** è fortemente impreciso e ambiguo, bene o male le parole contenute in un documento costituiscono i principali indizi dell'argomento in esso trattato.

## LA RICERCA SEMANTICA - 2

Migliorare la qualità della ricerca consiste in

- massimizzare i risultati rilevanti: **recall**
- minimizzare i risultati spuri: **precision**

L'uso delle keyword

- può migliorare la qualità della ricerca, specie se la classificazione dei contenuti è accurata e completa
- può orientare nell'effettuare la ricerca: si possono visualizzare indici inversi o grafici (distribuzioni) che evidenzino le keyword usate più di frequente.

E' però possibile raffinare la metodologia di classificazione e ricerca, e migliorane i risultati, introducendo l'uso di "strutture di conoscenza" come i **glossari**, i **thesauri**, le **ontologie**.

## GLOSSARI, THESAURI, ONTOLOGIE - 1

Un ***glossario*** è una raccolta di termini di un ambito specifico e circoscritto .. (Wikipedia)

- può fungere da riferimento terminologico per una pubblicazione o una collezione di documenti
- può servire a promuovere un linguaggio comune tra i membri di un'organizzazione o i partecipanti a un progetto.

Un ***thesaurus*** è caratterizzato di solito da

- uno "status" più ufficiale
- una struttura tassonomica: sono definite le relazioni *broader term* e *narrower term*
- la presenza di termini in più lingue
- l'uso di codici alfanumerici in aggiunta ai label lessicali dei termini.

## PRODOTTI PLONE PER GESTIRE GLOSSARI

Da anni esistono almeno due estensioni per Plone che consentono di creare e mantenere dei glossari.

### PloneGlossary

- è uno strumento più completo, un'estensione di Plone dedicata

<http://pypi.python.org/pypi/Products.PloneGlossary/1.4.0RC2>

### PloneHelpCenter

- è una suite di strumenti che supporta tutte le attività legate alla documentazione di un prodotto, in particolare di un software (come Plone stesso)
- tra i numerosi tipi di contenuto specializzati, PloneHelpCenter include i tipi **Glossary** e **Definition**.

<http://pypi.python.org/pypi/Products.PloneHelpCenter/3.0b3>

# PLONEGLOSSARY

From the PloneGlossary documentation:

## 1 Overview

PloneGlossary is a Plone content type that allows you to manage your own glossaries, propose definitions and search in one or more glossaries. Any word defined is instantly highlighted in the content of your site.

After adding a glossary, you can add your definitions to it. Definitions are a simple content type. Enter the word you want to define as the title, and the definition of the word in the text body. You can also specify variants of the word. For example if you define the word yoghurt, you may also want to allow the variants yogurt or yoghourt to be valid. Definitions will be highlighted (like an acronym) when they appear elsewhere in your site. (Also see the ploneglossary configlet.)

Once you have a large number of definitions in your glossary, you can browse the glossary by the means of an alphabetic index, or perform a search in the glossary. Each glossary has an integrated search engine, which is simply a ZCatalog.

## PLONEHELPCENTER

From the PloneHelpCenter documentation:

*A simple help-desk style documentation product for Plone.*

**Latest Version:** [3.0b3](#)

### Overview

Plone Help Center is an application designed to aid the documentation of Plone, and is used on plone.org to categorize and keep documentation up to date. It should be usable for documenting other open source products (such as Plone Product add-ons) or even for other documentation projects.

A ***glossary definition*** describes a particular term used as concisely as possible - typical definitions:

- ***CMF***: The Content Management Framework
- ***Workflow***: A state machine structure used to model business processes

## GLOSSARI, THESAURI, ONTOLOGIE - 2

In termini generali, un'**ontologia** è la "conoscenza condivisa di un dominio di interesse" (Usher).

Di solito essa si struttura come un insieme di **concetti** corredato dalle *definizioni* dei concetti stessi e dalle *interrelazioni* che sussistono tra i concetti e/o specifiche istanze dei concetti.

Possiamo anche dire che un'ontologia è una concettualizzazione di un dominio di interesse.

## GLOSSARI, THESAURI, ONTOLOGIE - 3

Spesso un'ontologia esplicita la **struttura tassonomica** del dominio di interesse, usando relazione di *generalizzazione / specializzazione* tra concetti.

Da questo punto di vista, un **thesaurus** assomiglia ad un'ontologia. Entrambi hanno una struttura tassonomica.

In un thesaurus si parla di *broader term* e *narrower term* (termine più generale e termine più specifico).

## CLASSIFICAZIONE E RICERCA BASATA SU ONTOLOGIE - 1

Quando si cerca di effettuare e supportare la **ricerca efficace ed efficiente** di pagine web, di documenti e di altri contenuti, non esistono confini netti tra i diversi approcci e le diverse tecniche.

La **ricerca per parole e frasi** in linea di principio può sembrare un po' rozza, ma in molti casi è quella che presenta il più alto rapporto prestazioni / costo.

Se però fossimo in grado di associare i documenti (e altri contenuti) ai concetti di un'ontologia o di un'altra struttura di conoscenza tassonomica, potremmo fare delle **ricerche semantiche** di indubbio interesse.

## CLASSIFICAZIONE E RICERCA BASATA SU ONTOLOGIE - 2

Per esempio potremmo trovare

- non solo in quali documenti si parla di "Fido" o "Luna" o "Bob"
- ma anche in quali documenti si parla di cani o in quali si parla di animali, anche se le parole "cane" o "animale" non sono presenti.

Analogamente potremmo cercare

- non solo in quali documenti si parla di alberghi, piscine, navi da crociera
- ma anche in quali documenti si parla di turismo o di vacanze, senza che questi termini figurino esplicitamente nel testo.

## ESTENSIONI DI PLONE PER THESAURI ED ONTOLOGIE

Per la gestione di thesauri e ontologie ho conoscenza di 3 prodotti:

- **PloneOntology**  
estende il meccanismo delle “keyword” nativo di Plone:  
"PloneOntology is an ontology based replacement for the existing keyword mechanism in Plone"  
<http://plone.org/products/ploneontology>
- **OWL Content**  
risultato della tesi di laurea preparata da uno studente dell'Università “La Sapienza” di Roma  
<http://www.hs01.it/area-comunicazione/press/press-releases/test>
- **PloneSaurus**  
sviluppato originariamente da LINK srl, per il progetto europeo INTEROP (una NoE con decine di partner).

## PLONEONTOLOGY - 1

### Funzionalità

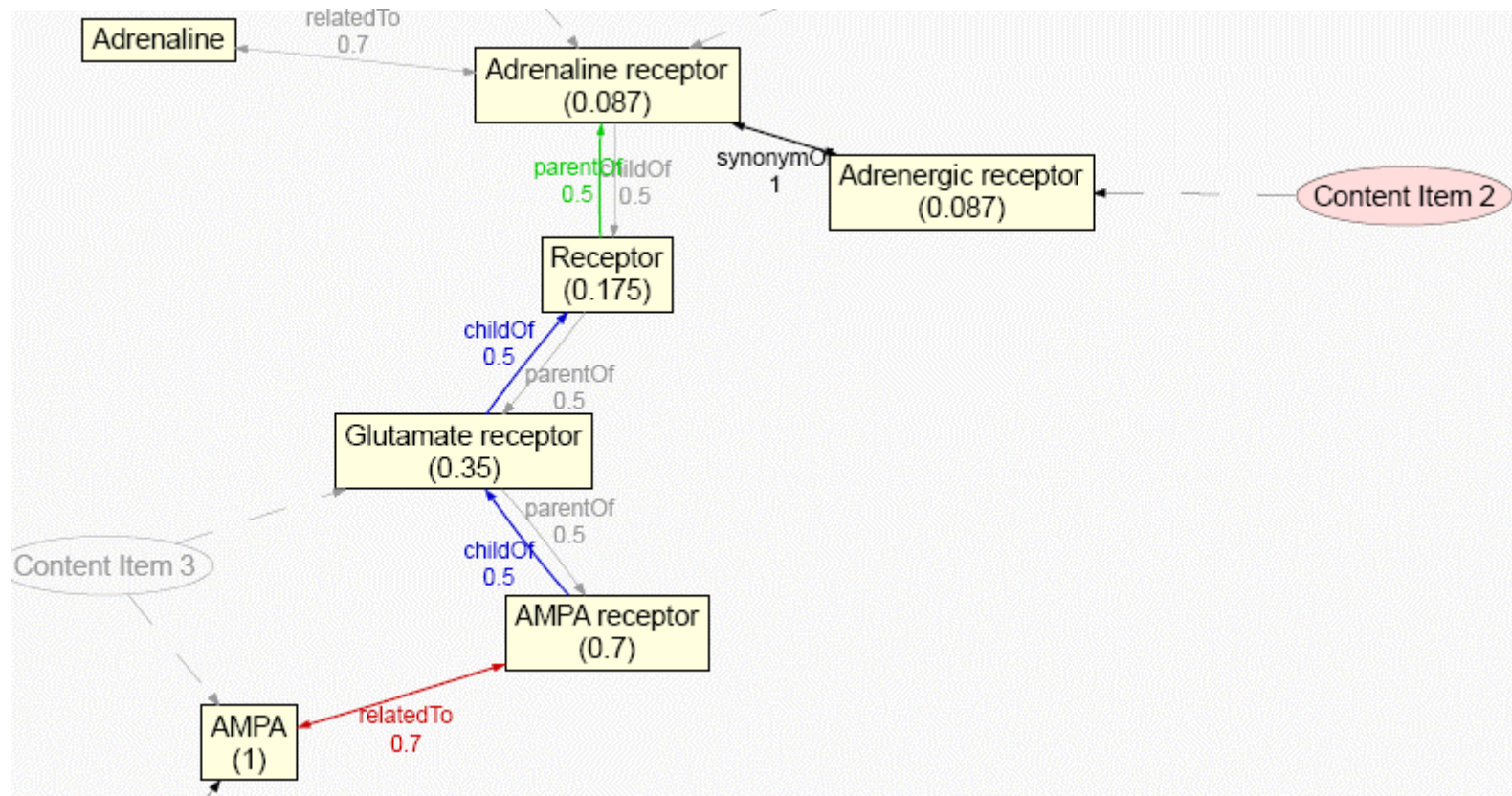
- i termini sono legati tra loro da relazioni arbitrarie
- i contenuti sono classificati relazionandoli ai termini
- visualizzazione grafica di tutte le relazioni
- la ricerca sfrutta relazioni tra termini e "pesi" associati
- creazione e manutenzione collaborativa dell'ontologia: gli utenti "propongono" nuovi termini e relazioni

### Alcuni problemi

- disponibile solo per Plone 2.1.1 e Plone 2.5
- non include visualizzazione ottimizzata per tassonomie
- i contenuti richiedono uno "schema" (Archetypes) esteso
- le "proposte" di nuovi termini e di nuove relazioni sono tipi di contenuto diversi dai termini e dalle relazioni "accepted".

## PLONEONTOLOGY - 2

Un esempio elaborato:



## OWL CONTENT

Dalla presentazione di Marco De Vitis:

OWL Content aggiunge al CMS la possibilità di inserire e visualizzare file OWL come normali contenuti.

Sviluppato seguendo gli standard, con uso di librerie esterne, installazione immediata

Tecnologie coinvolte: semantic web, web 2.0, CMS, XML, OWL, RDF, Python, XSLT

## PLONESAURUS - 1

PloneSaurus supporta la creazione di **tassonomie**, cioè di *glossari tassonomici* e *thesauri*; relazioni implementate:

- la relazione ISA: *generalizzazione / specializzazione*
- la relazione generica *related-to*

Sviluppato per Plone 2.1, è stato portato a

- Plone 2.5, Plone 3.1, Plone 3.5
- ma non abbiamo mai trovato il tempo per pubblicarlo; il problema principale è che nel tempo si sono stratificate troppe funzioni e troppi stili di interfaccia utente

Alcune caratteristiche

- varianti lessicali dei termini che “etichettano” un concetto
- definizioni multiple per i concetti
- possibilità di ristrutturare la tassonomia interattivamente.

## PLONESAURUS - 2

Il prodotto include

- un **consensus system**: i membri di un gruppo possono proporre e votare concetti e definizioni
- appositi *workflow* per gestire diverse fasi del ciclo di vita di una tassonomia
- funzioni di import/export da/a documenti OWL.

E' possibile

- creare diverse tassonomie in un sito Plone
- visualizzare graficamente le tassonomie, sia nel corpo della pagina, sia in portlet multi-tassonomia
- classificare un contenuto con termini da più tassonomie.

## PLONESAURUS - APPLICAZIONE KMAP - 1

PloneSaurus è stato sviluppato nell'ambito del progetto europeo **INTEROP**:


- il "consensus system" ha consentito di raffinare una tassonomia di 2000 termini
- classificazione e ricerca semantica dei contenuti della **KMap**: una "Knowledge Map" su attori, attività e risultati della ricerca europea nel campo della interoperabilità tra imprese e tra sistemi d'impresa.



Nell'applicazione KMap, PloneSaurus supporta classificazione e ricerca semantica dei contenuti:

- classificazione manuale; classificazione automatica, con "estrazione" dei termini-concetti dai documenti
- ricerca semantica, semplice o basata sulla nozione di *vicinanza semantica*.

## PLONESAURUS - APPLICAZIONE KMAP - 2

### Applicazione KMap - un termine della tassonomia sull'interoperabilità di impresa

 **model transformation**

▲ Up to glossary: *Last TAV Ontology*  



▲ Up to parent term: *transformation*

**Term label**  
model transformation


**Has variants**  
modeling transformation  
modelling transformation

**Flags**  
Competence domain, Enabling technology


**Term definition(s)**

1. Model transformation refers to the ability to establish a relationship between model elements or sets of model elements that represent the same concept in different models. 
2. Model transformation is the capability of describing relationships and mappings between model concepts (in effect metamodel elements) and thereafter executing the transformation on model instances. 

**Child terms (more specific concepts)**

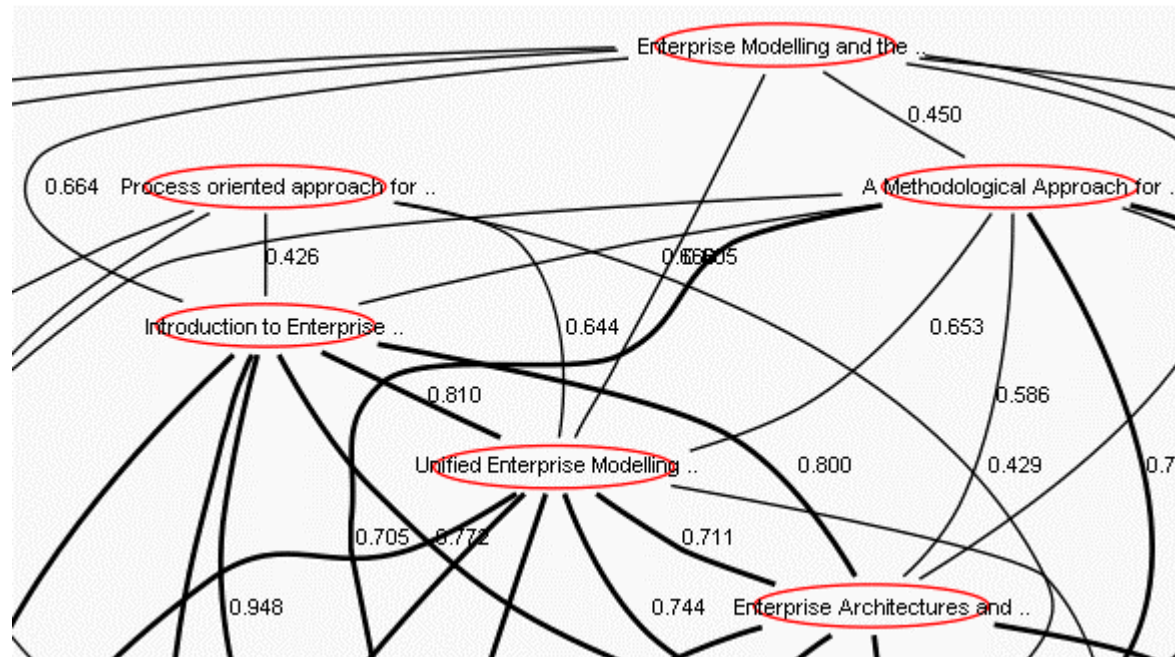
 [uml model transformation](#)

**Terms representing problems for which this term represents a solution**

 [interoperability barrier](#)

## PLONESAURUS - APPLICAZIONE KMAP - 3

Il risultato "clusterizzato" di una ricerca semantica di pubblicazioni: porzione della vista grafica.



## PLONESAURUS - APPLICAZIONE KMAP - 4

Classificazione automatica: *indice inverso* dei termini dopo la "batch annotation" di un lotto di documenti.

**List of the terms with their frequency after the analysis of 182 documents.**

*This is the list of the terms in the Taxonomy occurring in the documents analyzed.*

▲	term	terms count	doc. count	inv. doc. freq.
1	active model	3	3	0.89
2	activity cluster	1	1	1.00
3	activity diagram	27	17	0.67
4	actor model	4	3	0.89
5	adaptive application	1	1	1.00
6	adaptive application model	1	1	1.00
7	adaptive application monitoring	1	1	1.00
8	adaptive control	1	1	1.00
9	adaptive controller design	1	1	1.00
10	adaptive enterprise	1	1	1.00
11	adaptive enterprise architecture	3	2	0.93
12	adaptive learning technology	2	1	1.00
13	adaptive middleware platform	3	2	0.93
14	adaptive user interface	4	3	0.89
15	agent architecture	12	6	0.81
16	agent communication	7	7	0.79
17	agent design	13	5	0.83
18	agent interaction	4	3	0.89
19	agent model	6	4	0.86

## PLONESAURUS - ALTRI CASI D'USO

PloneSaurus è stato usato in molti altri progetti, tra cui

- l'ontologia di un altro, più piccolo progetto europeo
- un prototipo di "mappa" di esperienze e competenze, per un consorzio privato di imprese; il problema: disporre di informazione di facile accesso per rispondere a bandi di gara nel campo della sistemistica avanzata
- 2 repository digitali realizzati per l'ISPESL, Istituto Superiore per la Prevenzione e la Sicurezza del lavoro; il secondo è un piccolo repository pubblico:  
<http://ispesl-oeav.linkback.net>
- la "Competence Map" del progetto Cultura&Territorio del Dipartimento Patrimonio Culturale del CNR; rappresenterà l'offerta CNR nel campo delle tecnologie per la conservazione e valorizzazione del patrimonio culturale; include circa 10 tassonomie, alcune veramente estese!

<http://www.cultura-territorio.cnr.it>

## PLONESAURUS - CASI D'USO - ISPESL - 1

### ISPESL - Catalogazione semantica di **digital repository**

In entrambe le applicazioni realizzate per l'ISPESL, abbiamo implementato in Plone i 3 thesauri di riferimento per il dominio applicativo:

- CIS - Il thesaurus CIS è il principale strumento di indicizzazione dei documenti utilizzato dall'ILO/CIS Bulletin e dal database CISDOC
- EUOSHA-OSH - Vocabolario multilingue prodotto dall'Agenzia Europea e nuovo strumento di reference nel campo OSH (occupational safety and health)
- ATECO-NACE - Classificazione delle attività economiche, sviluppato in versione italiana dall'ISTAT e derivato dalla classificazione europea della CE.

## PLONESAURUS - CASI D'USO - ISPESL - 2

Un esempio di thesaurus nel sito dello "Osservatorio sui rischi domestici e negli altri ambienti di vita":

### EUOSHA-OSH



Vocabolario multilingue prodotto dall'Agenzia Europea e nuovo strumento di reference nel campo OSH.

reinizializza

I valori tra parentesi rappresentano il numero di termini più specifici.

Clicca sul segno + per vedere i termini più specifici associati ad un termine. Clicca sul segno - per nascondere i termini più specifici associati ad un termine.

OSH root

- ⊕ 00001A - Condizioni politiche, sociali ed economiche [6]
- ⊕ 08801A - Analisi e gestione del rischio [9]
- ⊖ 28521A - Rischi sul luogo di lavoro
  - ⊕ 28561B - Rischi biologici [5]
  - ⊕ 29241B - Rischi chimici [8]
  - ⊕ 39881B - Rischi relativi alla manipolazione di materiali [2]
  - ⊕ 40201B - Rischi multifattoriali sul posto di lavoro [1]
  - ⊕ 40281B - Rischi fisici [6]
  - ⊕ 41681B - Rischi derivanti da impianti, macchinari e attrezzature di lavoro [33]
  - ⊕ 45361B - Rischi psicologici ed organizzativi [13]
- ⊕ 46401A - Salute e infortuni [4]
- ⊕ 54161A - Lavoro e lavoratori [3]

# PLONESAURUS - CASI D'USO - CULTURA E TERRITORIO - 1

## CNR-DPC - Portale Cultura e Territorio

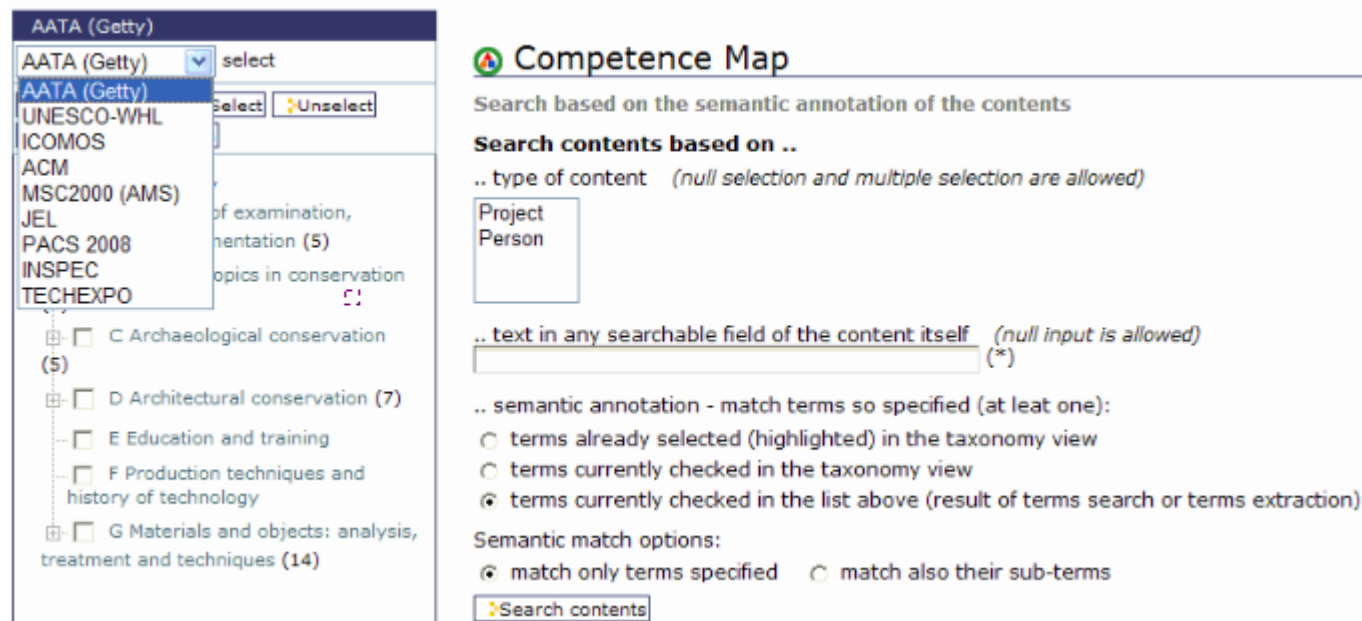
- la *Competence Map* implementa il modello concettuale del progetto; tutti i box e gli archi sono "attivi":



## PLONESAURUS - CASI D'USO - CULTURA E TERRITORIO - 2

### CNR-DPC - Portale Cultura e Territorio

- parte di un form di ricerca semantica, con portlet multi-tassonomia:



**AATA (Getty)**

AATA (Getty) select

AATA (Getty) select Unselect

UNESCO-WHL

ICOMOS

ACM

MSC2000 (AMS)

JEL of examination,

PACS 2008 mentation (5)

INSPEC topics in conservation

TECHEXPO

C Archaeological conservation (5)

D Architectural conservation (7)

E Education and training

F Production techniques and history of technology

G Materials and objects: analysis, treatment and techniques (14)

**Competence Map**

Search based on the semantic annotation of the contents

**Search contents based on ..**

.. type of content (null selection and multiple selection are allowed)

Project  
Person

.. text in any searchable field of the content itself (null input is allowed) (\*)

.. semantic annotation - match terms so specified (at least one):

terms already selected (highlighted) in the taxonomy view

terms currently checked in the taxonomy view

terms currently checked in the list above (result of terms search or terms extraction)

Semantic match options:

match only terms specified  match also their sub-terms

Search contents

## PLONESAURUS - CASI D'USO - CULTURA E TERRITORIO - 3

### CNR-DPC - Portale Cultura e Territorio

- parte di un form di ricerca: filtro testuale sui termini

Full-text search of terms

.. used in the annotation of some contents

.. whose label matches the text pattern below (use \* to specify null input if the option above is checked)

(\*)

.. only from the selected taxonomies (null selection and multiple selection are allowed)

AATA (Getty)  
 UNESCO-WHL  
 ICOMOS  
 ACM  
 MSC2000 (AMS)  
 JEL  
 PACS 2008  
 INSPEC  
 TECHEXPO

(\*) Please read the [help page](#) on full-text search

Result of search for terms matching the text in the box above

Last columns (click on header to sort): N=contents referring the term; N\*=contents referring the term or some subterms.

<input type="checkbox"/> Sel. All/None	Term ▲	Classification	N	N*
<input type="checkbox"/>	90.00.00 GEOPHYSICS, ASTRONOMY, AND ASTROPHYSICS 93.00.00 Geophysical observations, instrumentation, and techniques <b>93.30.-w Information related to geographical regions</b>	pacs	. 3.	. 3.
<input type="checkbox"/>	H. Information Systems H.2 DATABASE MANAGEMENT H.2.8 Database Applications <b>H.2.8.04 Spatial databases and GIS</b>	acm	. 4.	. 4.