

Extending the “Facets” concept by applying NLP tools to catalog records of scientific literature

*E. Picchi, *M. Sassi, **S. Biagioni, **S. Giannini

*Institute of Computational Linguistics

**Institute of Information Science and Technologies

CNR, Italy



The contest

National Research Council of Italy - CNR

Institute of Computational Linguistics

DBTficio Laboratory



Models and methods for the natural languages processing, monolingual and multilingual prototype applications

Institute of Information Science and Technologies

Networked Multimedia Information Systems Laboratory & Library



Digital Library management systems



The target

The target is to present the prototype of an “intelligent” navigation system named DBT&Facets, which has been implemented on the full bibliographic records of the documents archived in the PUMA digital library of the Italian National Research Council (CNR) <http://puma.isti.cnr.it>

The system has been implemented by integrating the core textual search engine (known as DBT and developed by ILC) with the TextPower (TP) technology.

<http://serverdbt.ilc.cnr.it/sitoDBT/>



PUMA repositories

PUMA is a user-focused, service-oriented infrastructure which manages an increasing number of CNR institutional repositories containing about 25,000 published or open access documents in a wide variety of disciplines

PUMA archives the metadata (qualified DC + administrative metadata) and the full texts of the following document types:

- Published literature: journal article, conference and workshop paper, book and contribution to book, guest editorial
- Grey Literature: conference presentation, workshop and meeting paper, communication poster/abstract, pre print, technical report, project report, internal note, PHD thesis, guest editorial, other materials (eg . courses, tutorials etc. ..)



TP-Text Power Technology

TP is based on NLP techniques and linguistic resources used to create tools for the evaluation, analysis, classification and browsing of information related to the domains of scientific literature

The extraction of implicit knowledge from the texts through which TP can enrich the documents, is a specialization of the "Facets" technology



The “facet” concept...

... is peculiar of Archives and Library Science field, but is also used in Information retrieval systems. In Library Science the term "facet" identifies the elements of a structured material such as library catalogs, which are characterized by the code of the field and its contents

TP extends the facet concept by extracting “field + content” pairs not only from structured fields but also from free text, eg. abstracts, using a linguistic-statistical approach to annotate relevant terminology, named entities, etc. The enriched text can be queried, analysed, and classified using “DBT&Facets”



...The prototype

DBT&Facets is an advanced search tool that permits the user to query and refine their results, and to identify particular relations between them



PUMA
Multidisciplinary
Repositories

Ca. 25000
Records &
Abstracts

enriched by



and elaborated by



“Intelligent” navigation system

The Puma query sample



Fielded search
[Default operator between fields is AND]

Abstract
Title
Subject
Author
Year

Select collection
"All Collection"
Start search Reset

Line x page: "ALL"

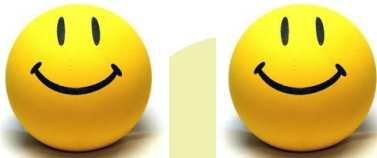
Simple search
[All Collection]
grey
Start search

N. 32 record(s) found

Selection from: 1 to: 32 ● Open access ● Restricted ● No access

● cnr.isti/2010-A2-004	Open Access to grey literature on e-Infrastructures: the BELIEF-II project digital library	Abstract language	English
● cnr.ilc/2010-A2-001	Grey Literature and Computational Linguistics: From Paper to Net	English abstract	Over the last ten years the impact of grey literature on conventional literature has frequently been studied. Studies have made use of bibliometric instruments used for citation analysis. Recently, this research has magnified attention on the impact of new forms of GL that have emerged along with the spread of Internet. This work aims to a) measure the impact of GL on two different scientific fields; b) describe the characteristics of GL documents cited; c) ascertain any changes in LG impact due to use of the www. Two years (1995 and 2003) were chosen as illustrative of the situation before and after the growth in the use of the www. With these aims, bibliographic references have been analysed in publications in two scientific fields for which it is logical to hypothesise a different impact. The publications are three journals of computer sciences included in the Journal Citation Report (JCR) Science Ed., and three journals of demography included in Journal Citation Report (JCR) - Social Science Ed.. The three journals in each of the two categories were chosen on the basis of their stability during the observation period (1995 and 2003) in terms both of their Impact Factor (IF) - high, medium and low - and of their ranks.
● cnr.isti/2009-A6-005	Open Access Library	Document language	English
● cnr.isti/2009-A3-013	BELIEF-II Pro	Subject(s)	Grey Literature
● cnr.istec/2009-A3-019	Archaeometri	Subject_ACM	A.1 Introductory and Survey Grey Literature
● cnr.irpps/2009-A6-002	From CNR Ar	Type	A2 International Conference
● cnr.ilc/2009-A6-001	Grey Literatur	Event Title	Work on Grey in Progress. Sixth International Conference on Grey Literature
● cnr.ilc/2009-A2-003	Grey Literatur Approach.		
● cnr.ieni.ge/2009-A0-015	Morphology a analysis		
● cnr.irpps/2008-A2-001	The impact of Scholar		
● cnr.ilc/2008-A6-001	Grey Literatur Approach		
● cnr.ibf.pi/2008-A0-002	An automatic speeds of swi		
● cnr.isti/2006-A2-03	Assisting scie experience be		
● cnr.isti/2006-A0-44	Assisting scie experience be		
● cnr.iqq/2006-B0-007	Petrological a Etba island, n		
● cnr.isti/2005-A3-23	Assisting scie experience be		
● cnr.isti/2005-A2-72	Object tracking in a stereo and infrared vision system		
● cnr.isti/2005-A2-24	Trend evaluation and comparison of the use and value of GL in core demography and computer science journals		

Simple Search for "grey"



The prototype query sample 1

Ricerca : grey [Faccette](#) [Ranking](#) [Chiudi](#) Trovati: 19

Faccette [Grafo faccette](#)



Ricerca Azzera

Cat.B: [cnr.iei](#) [cnr.ifc](#) [cnr.imati.ge](#) [cnr.irpps](#) [cnr.isti](#)

Cat.C: [andreoni, antonella](#) [baldacci, maria bruna \(nmis \)](#) [benvenuti, marco](#) [biagioni, stefania \(bib \)](#) [biagioni, stefania \(nmis \)](#) [biagioni, stefania](#) [biasotti, sylvia](#) [carlesi, carlo \(dir \)](#) [carlesi, carlo \(nmis \)](#) [carlesi, carlo](#) [castelli, donatella \(nmis \)](#) [colantonio, sara](#) [dattolo, p.](#) [de floriani, leila](#) [di bono, maria grazia](#) [di cesare, rosa](#) [falcidieno, bianca](#) [ferdeghini, ezio](#) [maria giannini, sylvia \(bib \)](#) [giannini, sylvia](#) [landini, luigi](#) [levi, g.](#) [lombardi, massimo](#) [luzi, daniela](#) [maggi, roberta](#) [maggiore, q.](#) [michelassi, claudio](#) [montanari, u.](#) [morales, maria aurora](#) [pagano, pasquale \(nmis \)](#) [pagano, pasquale](#) [papaleo, laura](#) [peters, carol \(nmis \)](#) [piacenti, mascia](#) [pieri, gabriele](#) [pisani, serena](#) [pizzarelli, f.](#) [positano, vincenzo](#) [romano, giuseppe alberto \(nmis \)](#) [ruggieri, roberta](#) [salvetti, ovidio](#) [santarelli, maria filomena](#)

Cat.D: [english](#)

Cat.F: [analisi delle citazioni](#) [analysis](#) [cardiac](#) [content-based retrieval](#) [contour tree](#) [digital libraries](#) [digital library](#) [document model](#) [etrdl](#) [grey literature](#) [image infrared vision](#) [introductory and survey.](#) [grey literature](#) [letteratura grigia](#) [morse complex](#) [morse theory](#) [motion prediction](#) [myocardial](#) [object tracking](#) [self publishing system](#) [self publishing](#) [shape description](#) [surface network](#) [uremics user guide](#)

Cat.G: [1970](#) [2003](#) [2005](#) [2006](#) [2007](#) [2008](#)

CATEGORIES

A	Collections
B	Institute
C	Author
D	Language of Summary
E	Language of Document
F	Free Subjects
G	Year of Publication



The prototype query sample 2

Ricerca : grey **Faccette** Ranking Chiudi **Trovati: 19**

Faccette [Grafo faccette](#)

Ricerca Azzera

libro con casa editrice internazionale comunicazioni/relazioni in convegni internazionali rapporti progetti di ricerca

Cat.I: a0 international journal a1 contribution to international book/monograph a2 international conference a3 international conference communication/abstract/poster a6 international conference abstract/poster pp pre print pr project report

Cat.J: french italian

Cat.K: d. j. farace and j frantzen d.j. farace and j. frantzen dominic farace dominic j. farace, j. frantzen

Cat.L: grey literature

Cat.M: conference on grey literature grey literature seventh international conference on grey literature

Cat.N: computer science digital library grey literature image segmentation reference digital library

Cat.O: access control applied mathematics artificial neural network cardiac magnetic resonance computer science data analysis digital library grey literature image segmentation imaging in vivo information science magnetic resonance reference digital library uremia

cnr.isti/2005-A2-24	1	the last ten years the impact of grey literature on conventional literature has frequently been studied. Studies have made use of bibliometric instruments used for citation analysis. Recently, this research has magnified attention on the impact of new ranks. Lingua documentò English Soggettò Grey Literature Soggettò_ACM A.1 Introductory and Survey. Grey Literature Tipo A2 International Conference Titolo evento Work on Grey in Progress. Sixth International Conference on Grey Literature Data evento 6-7 December 2004 Luogo evento Newü
cnr.irpps/2008-A2-001	2	cnr.irpps Titolo The impact of Grey Literature in the web environment: a citation analysis using Google Scholar Autore Di Cesare, Rosa Affiliazione IRPPS-CNR Autore Luzi, Daniela - Ruggieri, Roberta E-Mail r.dicesare@irpps.cnrü English Sommario in Inglese The use of Grey Literature (GL) has hitherto been studied onü Analisi delle citazioni Tipo A2 International Conference Titolo evento Grey foundations in information landscape. GL9 Ninth International conference on Grey literature Data evento 10-11 December 2007 Luogo evento Antwerpü

CATEGORIES	
H	Sub-Type
I	Type
J	Other Language of Summary
K	Edited by
L	Title of Journal
M	Title of Event
N	Title
O	Abstract in english



Conclusions

In an open domain like scientific documentation, our approach based on the criteria of “semantic similarity” is useful – and perhaps more objective than one based on hierarchical elements - as it makes it possible to link different types of information, also across domains if necessary

The aim of the project has been to structure a knowledge system of domain-specific information which assists the user by suggesting possible directions for their search