

Ricerca dell'informazione in rete - Parte seconda

Paola Zamperlin
3 maggio 2010

Laboratorio di Informatica. Corso di Laurea
in Lingue, letterature e studi interculturali
a.a. 2009-2010

Operatori booleani

Fondamentali per il recupero di informazioni in un archivio elettronico
Permettono di **combinare** più termini tra loro in una stessa interrogazione

AND

Trova soltanto i record che contengono tutti i termini inseriti nella stringa di ricerca

primo **and** secondo

OR

Trova i record che contengono sia:
- entrambi i termini inseriti nella stringa di ricerca,
- uno solo di essi

primo **or** secondo

NOT

Esclude i record con determinate caratteristiche

primo **not** secondo

Caratteri jolly

***** (asterisco): sostituisce un numero imprecisato di caratteri sia a destra che a sinistra del termine digitato
viagg* trova: viaggio, viaggi, viaggiatori ...

? (punto interrogativo): sostituisce un carattere all'interno del termine digitato
a?a trova ala, ara, aia ...

In Google

<http://www.google.it/support/websearch/bin/answer.py?answer=35889>

Google Books

<http://books.google.it/>

Progetto Gutenberg

http://www.gutenberg.org/wiki/Main_Page

Bartleby

<http://www.bartleby.com/>

Knovel

<http://www.knovel.com/web/portal/browse>

Leggere:

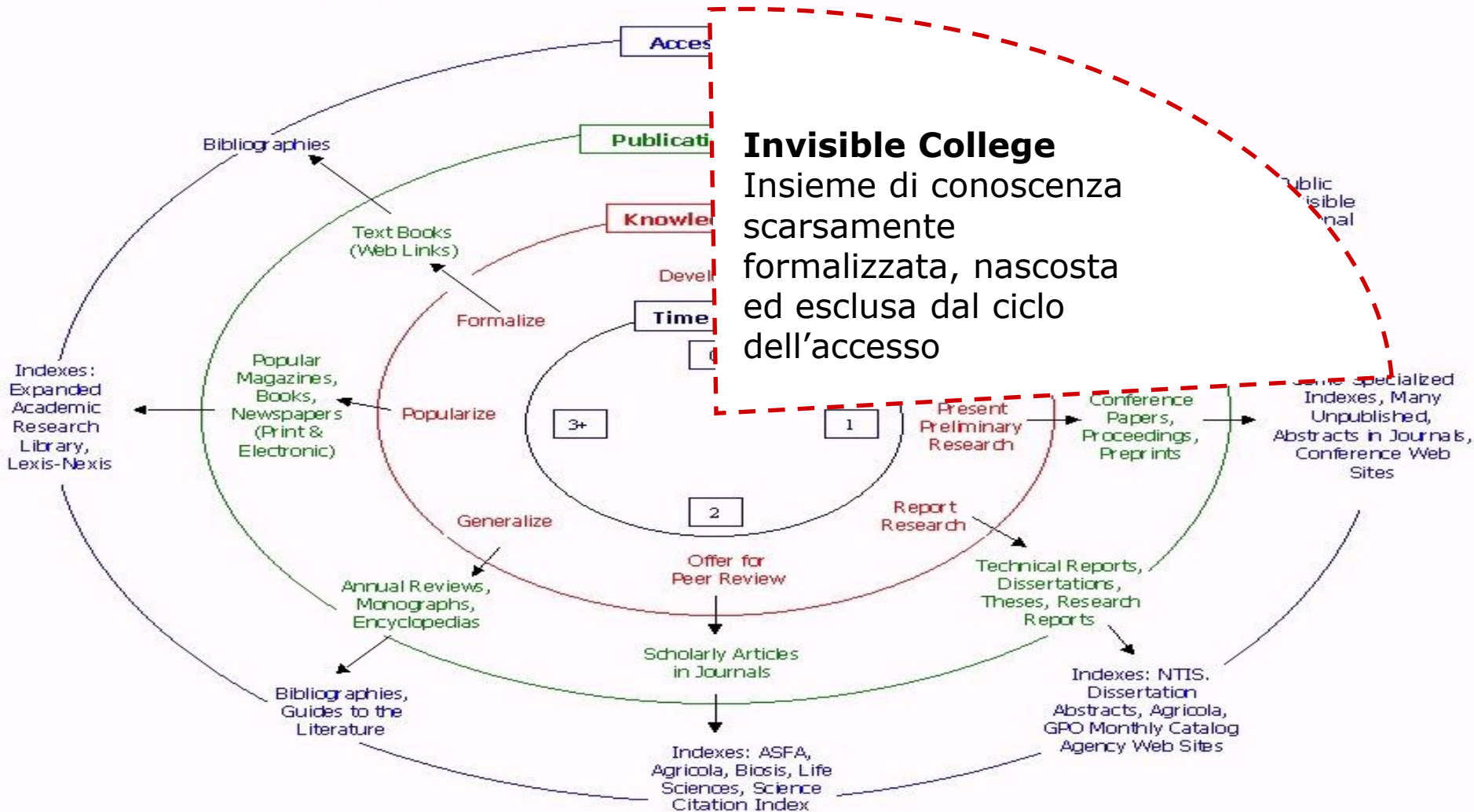
<http://oreilly.com/web2/archive/what-is-web-20.html>

Guardare:

<http://www.oreilynet.com/pub/e/1358>

Flusso della comunicazione scientifica - I

The Scientific Publication Cycle

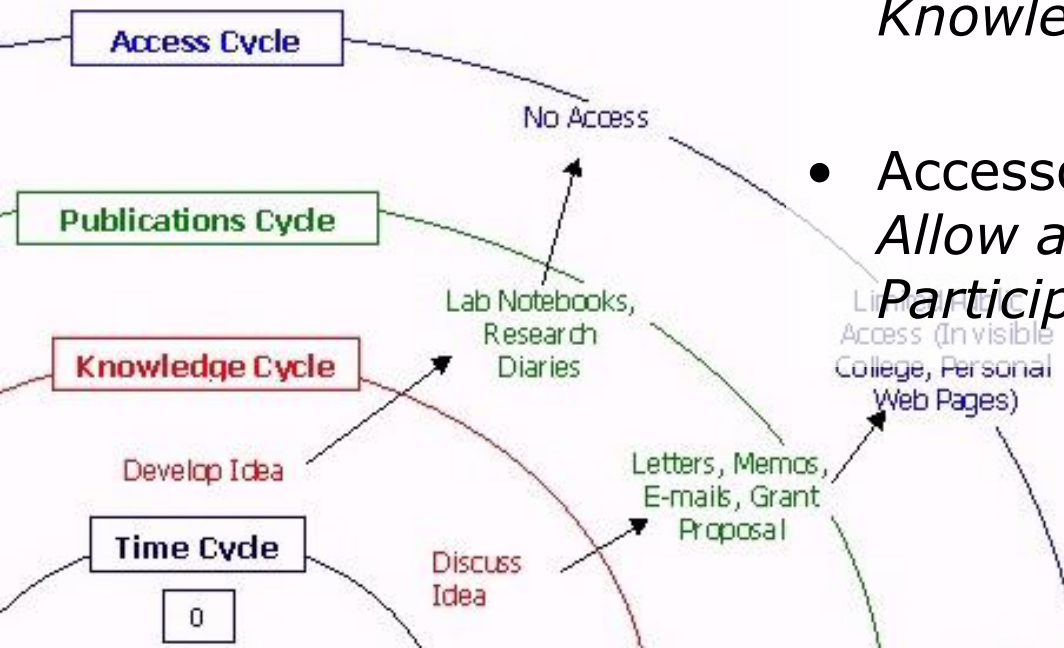


Flusso della comunicazione scientifica - II

Knowledge is conversation
(D. Lankes, 2007)

Le architetture informative si disegnano secondo una logica di

- Partecipazione [*People Need to be Active Constructor of their Knowledge*]
- Accesso [*People Want Tools that Allow and Facilitate Conversation and Participation*]



La comunicazione scientifica cambia

- Crescente quantità di **letteratura "grigia"** (testi preparatori o incompleti, set di dati, report, proposte di ricerca, diari di laboratorio) e non peer reviewed (preprint). Questa tendenza si accentua col cd Web 2.0: "scrivere, non solo leggere";
- Ruolo della **conoscenza tacita** che si esprime nell'attività collaborativa e di condivisione degli strumenti del Web 2.0

Dichiarazioni di Berlino e di Messina

4-5 novembre 2004

Per la prima volta nella storia, Internet offre oggi l'occasione di costituire un'istanza globale ed interattiva della conoscenza umana e dell'eredità culturale e di offrire la garanzia di un accesso universale ...

L'autore(i) ed il detentore(i) dei diritti relativi a tale contributo garantiscono a tutti gli utilizzatori il diritto d'accesso garantito, irrevocabile ed universale e l'autorizzazione a riprodurlo, utilizzarlo, distribuirlo, trasmetterlo e mostrarlo pubblicamente ...

Una versione completa del contributo e di tutti i materiali che lo corredano, inclusa una copia della autorizzazione come sopra indicato, in un formato elettronico secondo uno standard appropriato, è depositata (e dunque pubblicata) in almeno un archivio in linea che impieghi standard tecnici adeguati (come le definizioni degli *Open Archives*) e che sia supportato e mantenuto da un'istituzione accademica, una società scientifica, un'agenzia governativa o ogni altra organizzazione riconosciuta che persegua gli obiettivi dell'accesso aperto, della distribuzione illimitata, dell'interoperabilità e dell'archiviazione a lungo termine

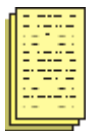
http://oa.mpg.de/openaccess-berlin/BerlinDeclaration_it.pdf

<http://www.aepic.it/conf/viewpaper.php?id=49&cf=1>

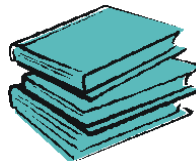
Impact cycle begins:
Research is done



Researchers write pre-refereeing
"Pre-Print"



Submitted to Journal



Pre-Print reviewed by Peer Experts – "Peer-Review"



Pre-Print revised by article's Authors

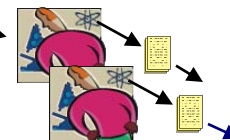
Refereed "Post-Print" Accepted, Certified, Published by Journal



Researchers can access the Post-Print if their university has a subscription to the Journal



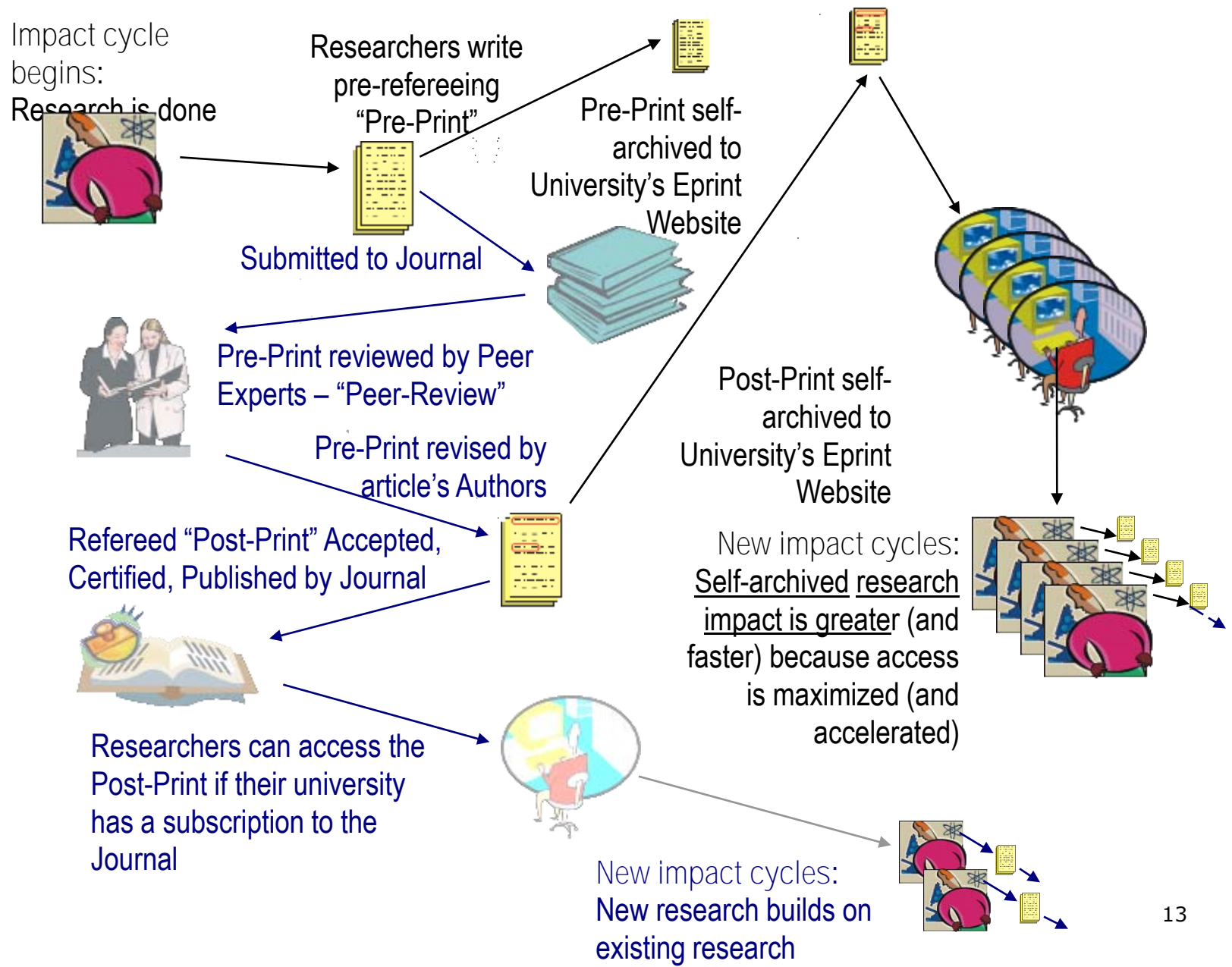
New impact cycles:
New research builds on existing research



12-18 Months

Maximized Research Access and Impact Through Self-Archiving

12-18 Months



I repositories di preprint e postprint

L'analisi della catena del valore delle pubblicazioni scientifiche, evidenzia il paradosso che le università finanziano la ricerca e stipendiano i ricercatori, ma devono poi ricomprare i risultati della ricerca dagli editori. È questa una delle cause del movimento Open Access che ha dato vita al programma **Open Archive Initiative (OAI)**.

Due strategie dell'open access initiative

Archiviazione in depositi disciplinari o istituzionali: e-prints di Firenze <http://e-prints.unifi.it/>

Creazione di riviste scientifiche ad accesso libero: DOAJ
<http://www.doaj.org/>

Harvesting e ricerca sui metadati
OAIster

<http://www.oaister.org/>

PLEIADI (Portale per la Letteratura scientifica Elettronica Italiana su Archivi aperti e Depositi Istituzionali) scaturisce dalla collaborazione fra due importanti consorzi interuniversitari italiani, CASPUR e CILEA, nell'ambito del progetto AEPIC

<http://www.openarchives.it/pleiadi/modules/openarchives/>

Tipi di documenti

Appunti

Mail

Set di dati

Dispense

Tesi

Interventi a incontri scientifici

Buone pratiche

Documenti preparatori

Preprint e postprint

...

Altri strumenti per la ricerca

Emergono nuovi modelli organizzativi che permettono accesso alle risorse e a numerose informazioni aggiuntive:

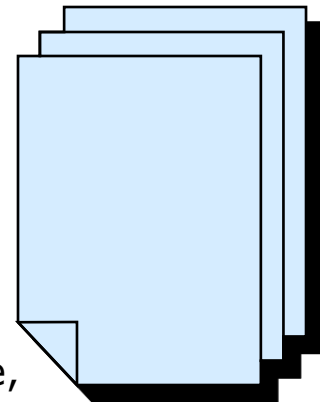
Amazon.com

(Immagine della copertina; Estratti dal libro; Search inside; Valutazione (1-5); SIPs (statistical improbable Phrases); CAPs (persone, luoghi, eventi, argomenti molto citati nel libro); Raccomandazioni basate sul comportamento di altri utenti che hanno comprato il libro; Consigli di libri simili; Personalizzazione tramite profilo utente; Indica altri fornitori che posseggono il libro cercato; Analisi frequenza delle parole contenute (concordance); Valutazione di leggibilità)

Books.google.com

Scholar.google.com

(ricerca sulla letteratura accademica: documenti approvati per la pubblicazione, tesi, libri, abstract e articoli di case editrici accademiche, ordini professionali, database di studi non ancora pubblicati, università e altre organizzazioni accademiche)



Incidenza del criterio di rilevanza e popolarità

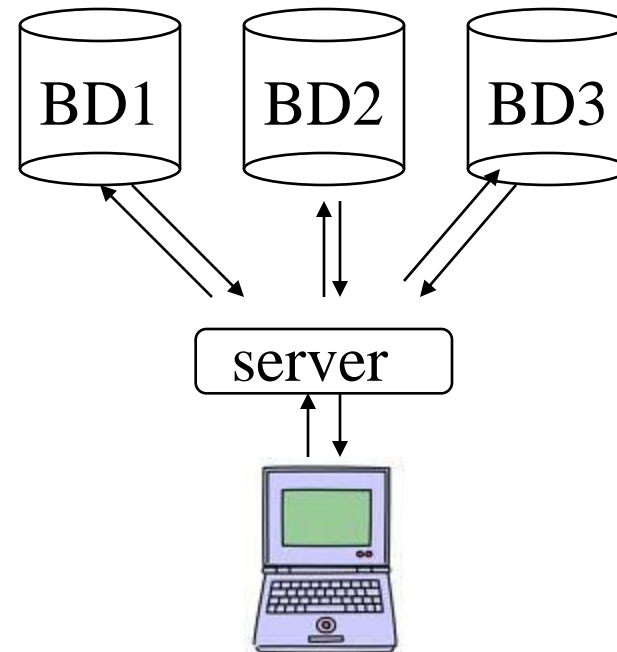
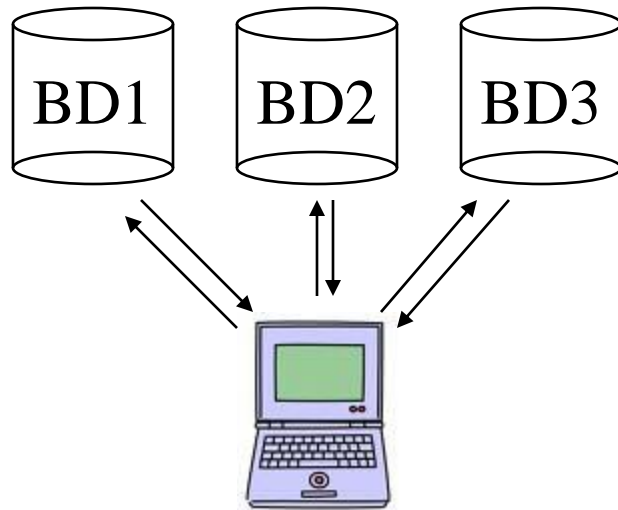
Banche dati

Generalmente nelle banche dati è più curata che nei cataloghi bibliotecari la ricerca in generale e in particolare quella semantica con l'aiuto di thesauri e di classificazioni. Tuttavia:

1. le banche dati sono costituite sulla base di progetti editoriali specifici e indipendenti, mancano di completezza e spesso si sovrappongono nella copertura disciplinare
2. la banca dati fornisce solo l'informazione e non la collocazione del documento: l'articolo deve essere ricercato altrove

Metasearch

Per ovviare al primo problema, la necessità si consultare numerose banche dati per raggiungere una affidabile copertura della ricerca, si applicano software di metaricerca

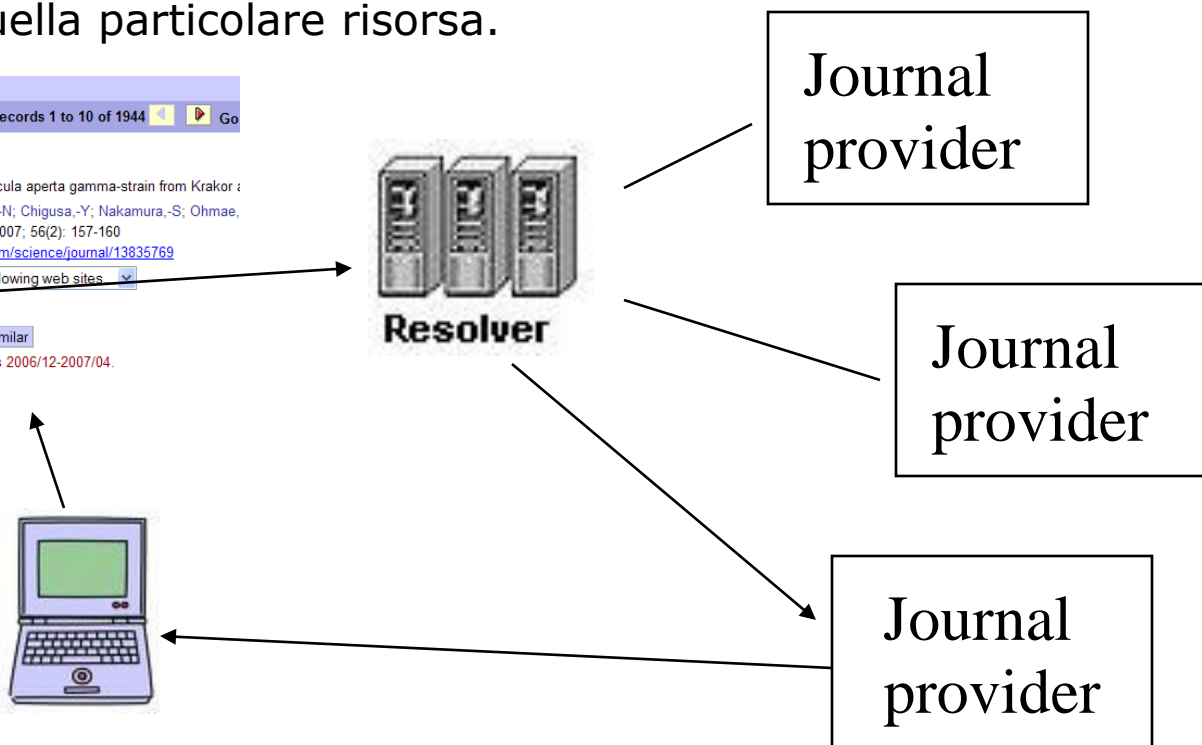
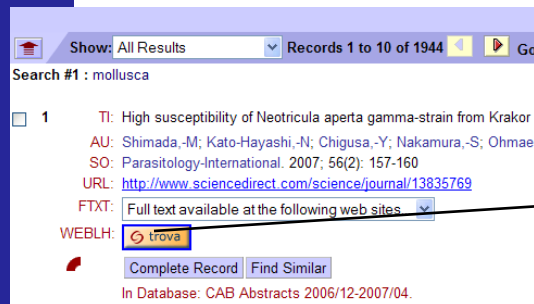


OpenURL

Per ovviare al secondo problema, si adotta il software OpenURL che permette di arrivare al full text di un articolo partendo da una citazione

L'OpenURL è un protocollo Internet per l'interoperabilità tra le risorse.

Il server SFX accetta i metadati dall'OpenURL, li analizza, intercetta l'identità dell'utente e crea dinamicamente i collegamenti ai servizi legati a quella particolare risorsa.



Strumenti per la ricerca terminologia

Tesauri disciplinari:

American Society of Indexers:

<http://www.asindexing.org/site/refbooks.shtml#science>

Web Thesaurus Compendium:

<http://www.ipsi.fraunhofer.de/~lutes/thesoecd.html>

Ricerca per citazioni

Importante servizio per accrescere la conoscenza contenuta negli archivi aperti è costituito dal *citation linking* : la possibilità di seguire le citazioni che vengono fatte o ricevute dai singoli documenti in modo da ricostruire il percorso intellettuale di una scoperta o di valutare l'impatto di un articolo su un argomento.

Il modello è Citeseer

<http://citeseer.ist.psu.edu/>

- ó Rivela relazioni con altri articoli (in base a frasi comuni e a co-citazioni)
- ó Rende conto di correzioni o smentite dei risultati
- ó Indica l'impatto di ciascun articolo
- ó Evita la ripetizione di ricerche già fatte
- ó Permette l'analisi di trend di ricerca e aiuta a identificare aree scientifiche emergenti (grafo delle citazioni)
- ó Evidenzia dove e quante volte un articolo è citato
- ó Possibilità di esprimere un giudizio secondo una scala da 1 a 5
- ó Possibilità di scrivere commenti

<http://www.citebase.org/>

Copyright di Tim Brody, University of Southampton. In fase di sperimentazione.

Facendo riferimento al solo microcosmo dei documenti contenuti nei depositi utilizzati, indica:

- Il valore d'impatto del documento in base a:

É le citazioni ricevute da altri articoli

É il numero di volte (hits) in cui il documento è stato scaricato (full text download) (basato solo su arXiv.org dal 1999)

- Il valore d'impatto di un autore:

É il numero totale di citazioni a documenti in cui è nominato l'autore, diviso per il numero di documenti in cui lo stesso autore è nominato.

É Il numero totale di volte (hits) in cui i documenti che citano un determinato autore sono stati scaricati (v. supra), diviso il numero dei documenti che citano lo stesso autore

É Citation/hits history: grafico delle due grandezze: numero/tempo

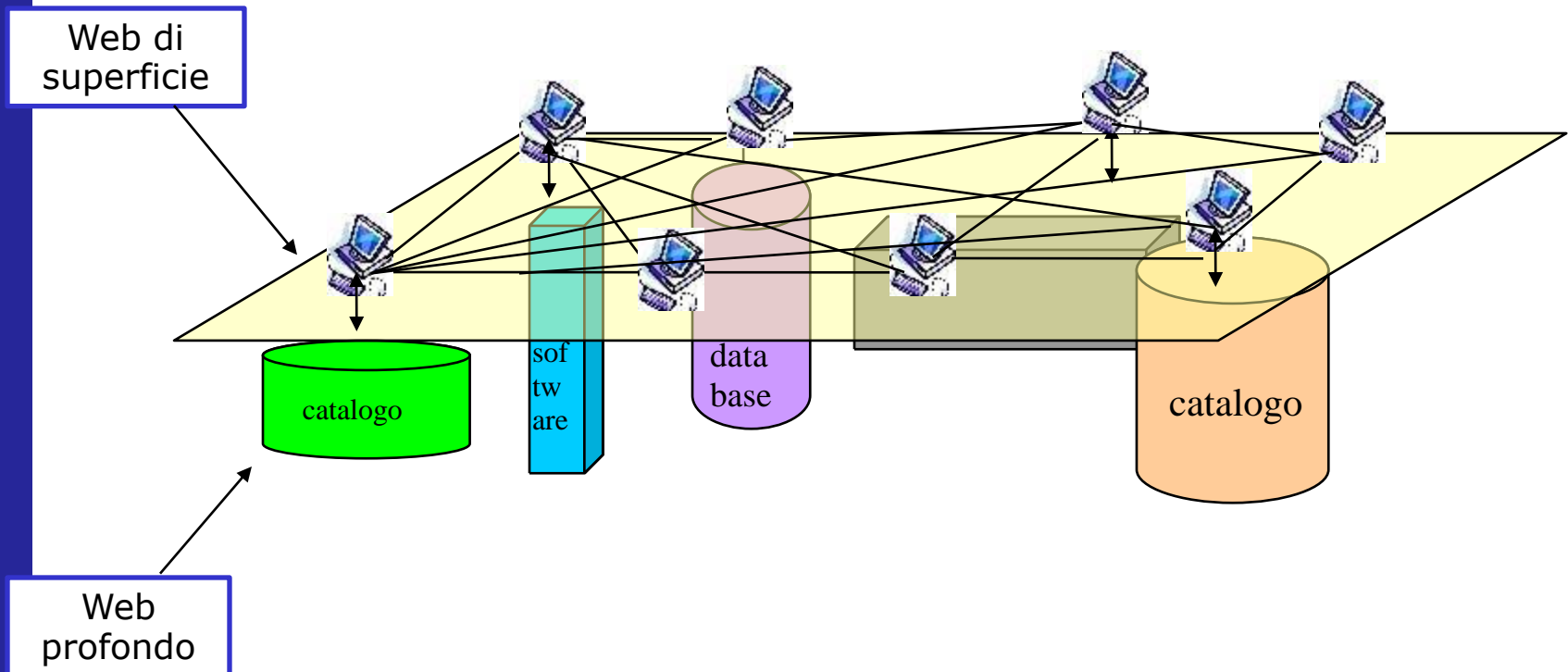
É valutazione di autorevolezza secondo una metrica sperimentale

É correla gli articoli in base alla citazione degli stessi documenti

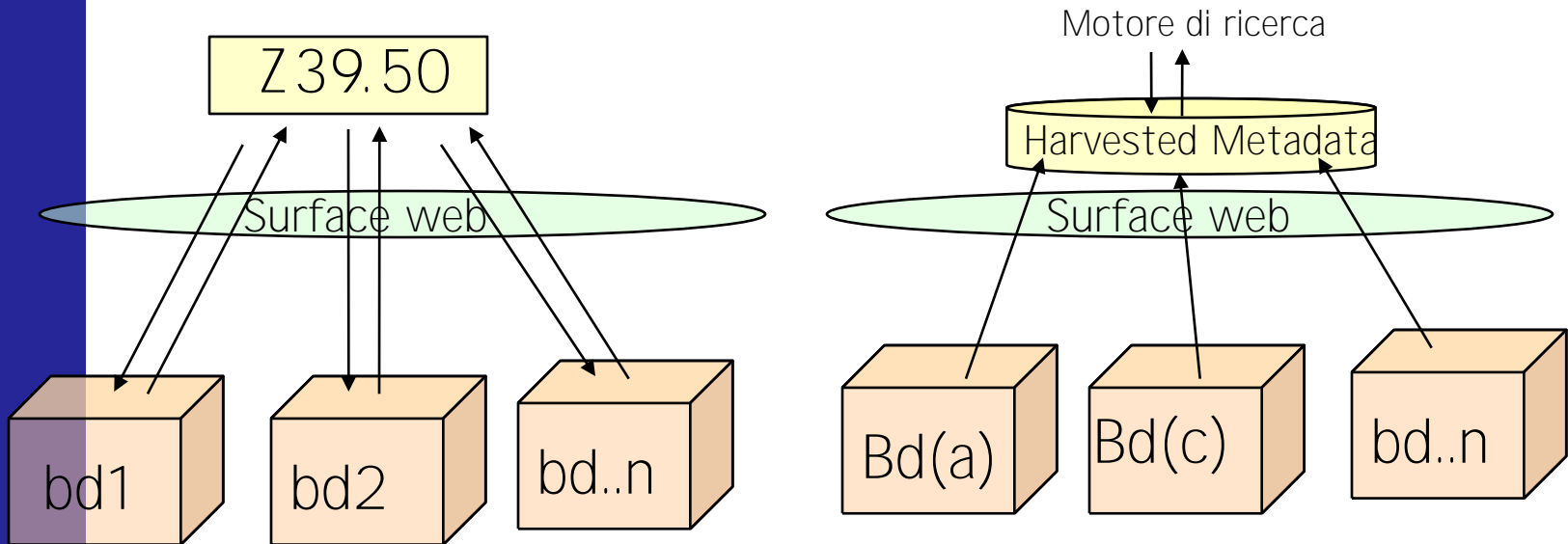
Il Web profondo

La grande discriminante per la costruzione di servizi avanzati, dipende sempre più:

- ó dallo stare sul web o fuori dal web
- ó dall'utilizzare standard riconosciuti e non proprietari
- ó dalla adozione dei cosiddetti "Web services"



La ricerca nel Web profondo



Z39.50 invia in sequenza la query alle singole basi dati seguendo le caratteristiche di ricerca offerte ed evidenzia i risultati in successione.

I metadati vengono **"esposti"**, raccolti (harvested) da un service provider (es. oaister) e la ricerca avviene nel deposito di metadati

I Web services

Un Web service è un software per permettere l'interoperabilità tra macchine nella rete

I Web services usano standard e protocolli aperti. I protocolli e i formati dei dati sono scritti in formato testo (utilizzando il metalinguaggio XML) rendendone facile la comprensione per gli sviluppatori

I Web services permettono di combinare facilmente software e servizi di differenti compagnie per costruire servizi integrati

Applicazioni software scritte in vari linguaggi di programmazione e che funzionano su diverse piattaforme possono usare i web services per scambiarsi dati sulle reti di computer in una maniera simile a quanto avviene in un singolo computer
(<http://www.wikipedia.org>)

Motori di ricerca di Internet filtering

Cosa si cerca: documenti testuali (html, txt, pdf, doc, etc.),
immagini, video, audio, etc.

Criteri di ricerca

Per parole:

operatori logici: and, or, not, and not

operatori vari: near, "", +, -

Operatori di quantità: <, >, =, #

(JPEG(7))

Criteri di rilevanza dei risultati

Frequenza diretta delle parole cercate

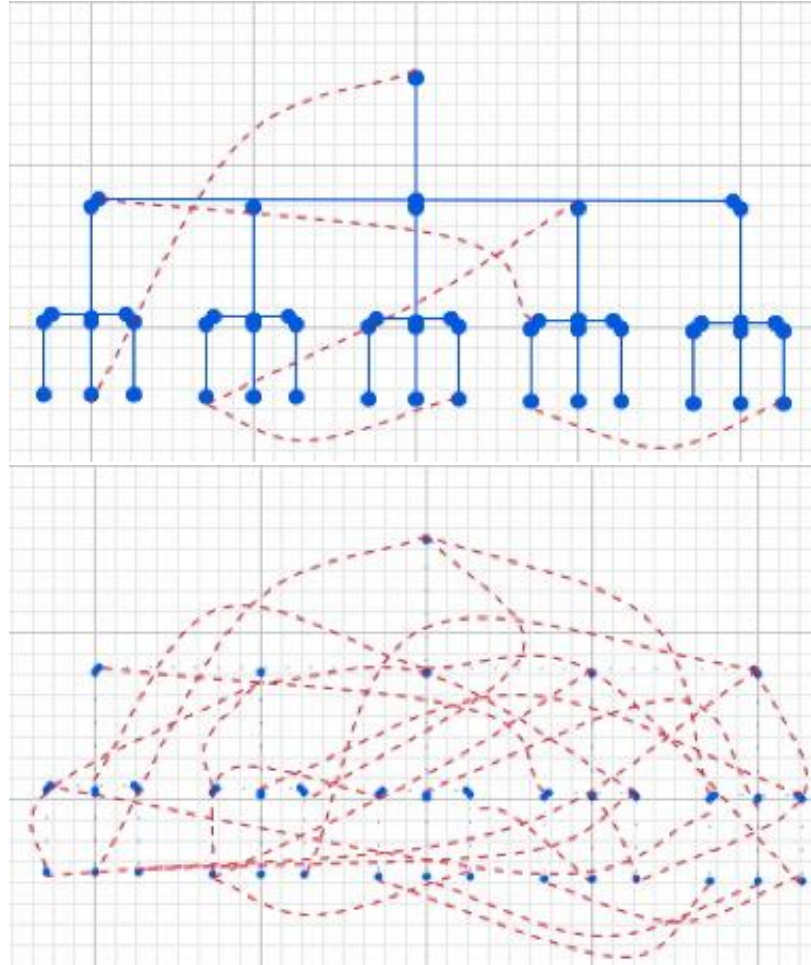
Frequenza inversa delle parole cercate

Posizione delle parole nella struttura del documento (nel
titolo, nella formattazione di intestazione, nella
descrizione di link, etc.)

Popolarità

Categorializzazione

Dalla ricerca gerarchica al filtering



Dalla precisione della ricerca alla serendipity

1. La produzione di conoscenza è sempre più interdisciplinare. Ciò significa che esistono sempre maggiori discontinuità nella organizzazione disciplinare del sapere che determinano una riconsiderazione delle modalità stesse della ricerca dell'informazione
2. Alla precisione della ricerca si aggiunge il meccanismo della *serendipity*, ad esempio nel social tagging, tramite la navigazione tra i tag assegnati ad una stessa risorsa da utenti posti in punti di osservazione diversi

Panorama dell'informazione

Col web si pubblica con grande facilità "di più con meno"
(esperienze conoscitive più ricche, grazie alla
multimedialità, con meno sforzo)

I testi, sia libri che riviste, sono sempre più disponibili in
formato elettronico, spesso gratuito e sempre più si
profila per il settore scientifico la prospettiva dell'Open
Access

Si ha una produzione di informazione rilevante e affidabile
che non segue i canali istituzionali né la tradizionale
"catena del valore" che fa riferimento all'editore
tradizionale

Al tradizionale meccanismo del peer review si aggiunge
quello più sfilacciato, ma in molti contesti efficace, del
"controllo" della comunità di interessi

Moltiplicarsi delle tipologie e dei formati delle risorse
informative.

Dall'information overload al metadata overload

Mario Rotta, *Content is dead, long live the content*

Metadata overload

Crescente ricorso a metadati automatici.
Crescente protagonismo degli utenti anche nell'organizzazione della conoscenza ("Al dilettantismo della pubblicazione non può non corrispondere un dilettantismo della catalogazione", Clay Shirky).

Ma il problema è solo una questione di precisione e coerenza nella assegnazione di metadati?

Tendenze

- É Dalla ricerca strutturata alla ricerca per parole libere
- É Dalla precisione della ricerca alla ricerca per rilevanza
- É Valorizzazione dei metadati impliciti: l'algoritmo della popolarità
- É Portare l'informazione delle biblioteche fin dove l'utente ne ha bisogno tramite il web: dal record ai metadati (OCLC: Find in a library); portare l'utente dall'informazione di Internet alla biblioteca (OpenURL Referrer; trova@unifi.it)
- É Partecipazione dell'utente (commenti, recensioni, folksonomie)

Partecipazione: l'utente della rete non è passivo bensì crea e condivide i propri contenuti sulla rete

Personalizzazione: organizzando le proprie risorse sulla rete l'utente crea valore aggiunto ai servizi che utilizza

Mashup: nuovi servizi si creano dalla aggregazione e ricombinazione di servizi preesistenti grazie ai Web services

Il ruolo dell'utente finale

Utilizzando i servizi di rete per organizzare le proprie risorse, l'utente finale conferisce valore aggiunto al servizio a beneficio di tutti

L'utente crea inoltre metadati in modo implicito:

- ó nel Web visitando i siti (criterio della popolarità)
- ó tramite profili utente che alimentano sistemi di raccomandazione delle risorse (cfr. Amazon)
- ó nell'attività di condivisione/collaborazione

Multimedialità

Video, sonoro, immagini richiedono una banda di comunicazione molto ampia che difficilmente un normale utente delle rete può permettersi; esistono però siti dove è possibile distribuire questi materiali appoggiandosi a servizi gratuiti

[Flickr](#) ; [Picasa](#) condivisione di foto;

[Youtube](#) ; [Google video](#); condivisione di video

[Slideshare](#) ; condivisione di presentazioni in slides

La disintermediazione: social tagging

Importante non è "chi categorizza meglio di me", bensì "chi categorizza come me": è importante sapere chi ha fatto un certo tag a una risorsa (cioè quale profilo di interessi ha) e quando (quanto è aggiornato il tag)

Il collegamento tra schemi individuali non avviene solo a livello di tag uguali, ma anche tramite i documenti conosciuti e identificati tramite URL: il passaggio è dal contenuto verso i punti di vista (tag)

La corrispondenza tra tag (es. sinonimia) non è binaria (si/no) ma analogica, cioè prefigura sovrapposizione parziale di domini e quindi implica diversità di contenuto

Socializzare e condividere

Social bookmarking: *users, tags, and URLs*

Del.icio.us: <http://delicious.com/> condivisione di bookmarks

Connotea: <http://www.connotea.org/> lanciato dal Nature Publishing Group per la condivisione di bookmark in ambiente di ricerca (v. D-Lib Magazine, April 2005).

Citeulike: <http://www.citeulike.org/> espressamente rivolto al mondo accademico. Si può costruire una bibliografia di documenti di rete semplicemente cliccando su un bookmarklet: se la citazione è tratta da alcuni siti predefiniti (tra cui molti archivi aperti e banche dati) il bookmark acquisisce direttamente i metadati del documento; altrimenti la maggior parte dei campi deve essere riempita a mano tramite un form abbastanza dettagliato.

Inserimento di riferimenti a risorse di rete tramite

É bookmarklet che aprono una pop up windows con il form dei metadati che si possono inserire; con alcuni siti i metadati sono aggiunti automaticamente

É Il comando Add form

É Indicando il DOI

Possibilità di indicare un OpenUrl Server per verificare la disponibilità di full text di citazioni

Possibilità di copiare nel proprio folder indicazioni inserite dal altri

Utilizzo di tag per designare il contenuto della risorsa (*volksonomie*);

É condivisione di tag

É Possibilità di specificare l'ambito semantico con *tag note*

Possibilità di aggiungere *commenti* alla risorsa e di visualizzare i commenti di altri

Folder individuali o di gruppo

Importazione esportazione di bookmark (formati RIS e Firefox, presto anche in formato BibTeX; è anche prevista la possibilità di importare bookmark da del.icio.us or CiteULike)