online information 2006

Sesión Española
Londres, 29 de Noviembre de 2006

Thinkepi

Bibliotecas, repositorios y otras bases de datos Web

Isidro F. Aguillo
CINDOC-CSIC

InternetLab
Observatory of Science & Technology in Internet

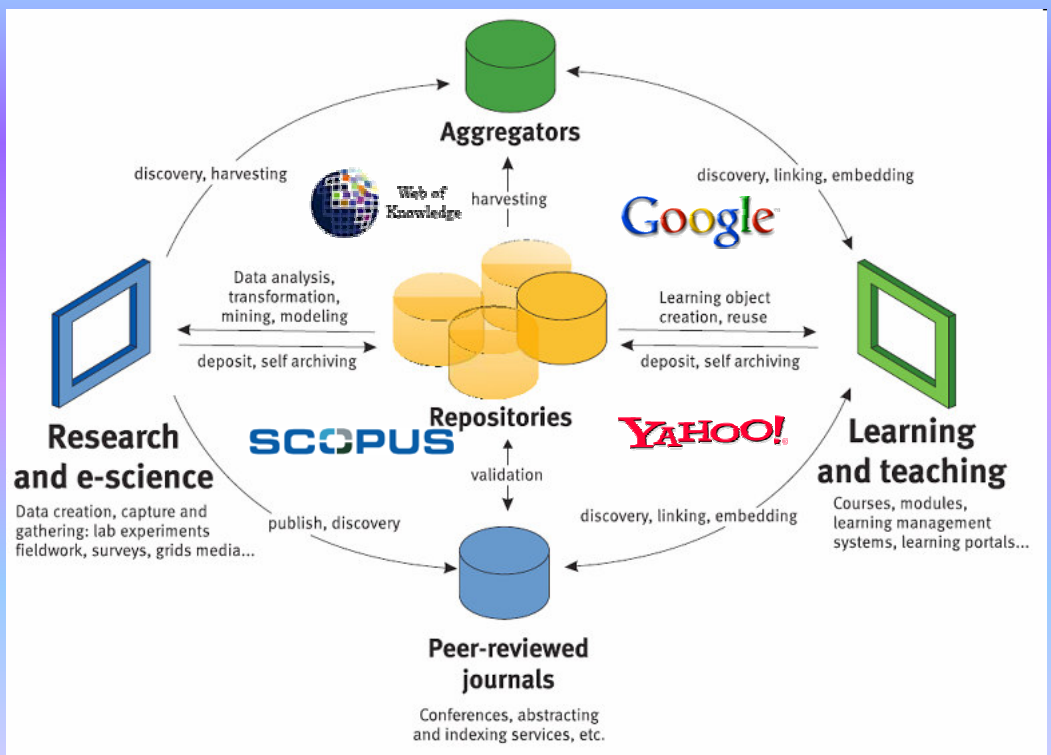
Una nueva realidad …

- Bibliotecas digitales "virtuales"

  Colecciones de documentos distribuidas

  - **Web visible**: Auto-archivo, depósitos de documentos, buscadores de area
  - **Web invisible**: Bases de datos, repositorios institucionales y temáticos, revistas electrónicas
  - **Internet privada**: Consorcios nacionales, regionales e institucionales

… para un nuevo modelo



Tres escenarios

- Web pública
  - Web visible
    Información recogida en los motores de búsqueda
  - Web invisible
    Bases de datos bajo pasarelas no indizadas por los motores de búsqueda
- Internet privada
  Se requiere subscripción o identificación vía IP

Recursos académicos en el Ciberespacio

Web Pública

Internet Invisible

Web Visible

Google

YAHOO!

msn Search

Ask.com

Bases de datos

Google Scholar

SCIRUS

Repositorios

arXiv.org

OAIster

RECERCAT

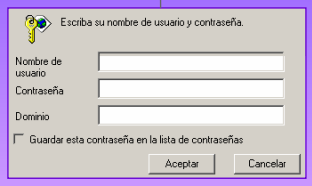
Revistas electrónicas

HighWire Press

PLoS

SciELO

Web Privada



Biblioteca Virtual del CSIC

SCOPUS

Web of Knowledge

ELSEVIER

Análisis cuantitativo (I)

- Web visible

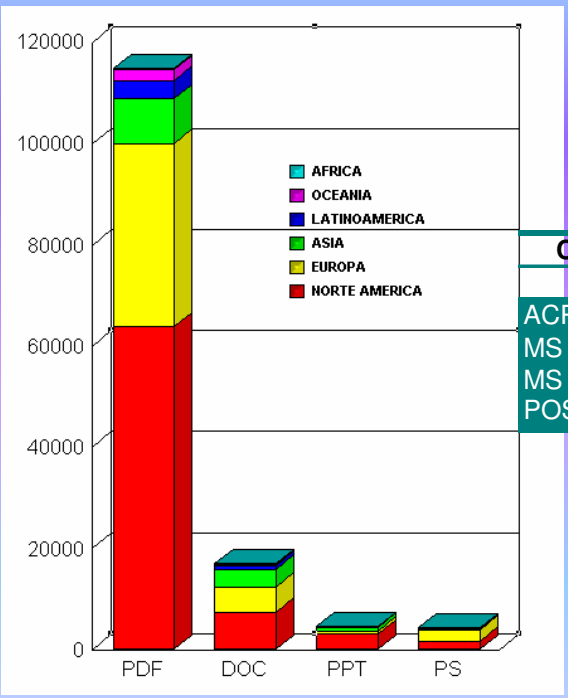
  Objetos: Ficheros ricos
  (autoarchivo, páginas personales e institucionales, depósitos)

  - Adobe Acrobat (.pdf): Artículos, informes, monografías
  - MS Word (.doc): material definitivo y borradores
  - Postscript (.ps): Artículos (Física, Matemática, Ingeniería, …)
  - MS Powerpoint (.ppt): congresos, docencia

- Método de acceso: Motores de búsqueda

  Rápido y flexible
  Alta visibilidad e impacto (acceso gratuito?)
  Sin control documental
  Cobertura escasa e irregular

Ficheros ricos



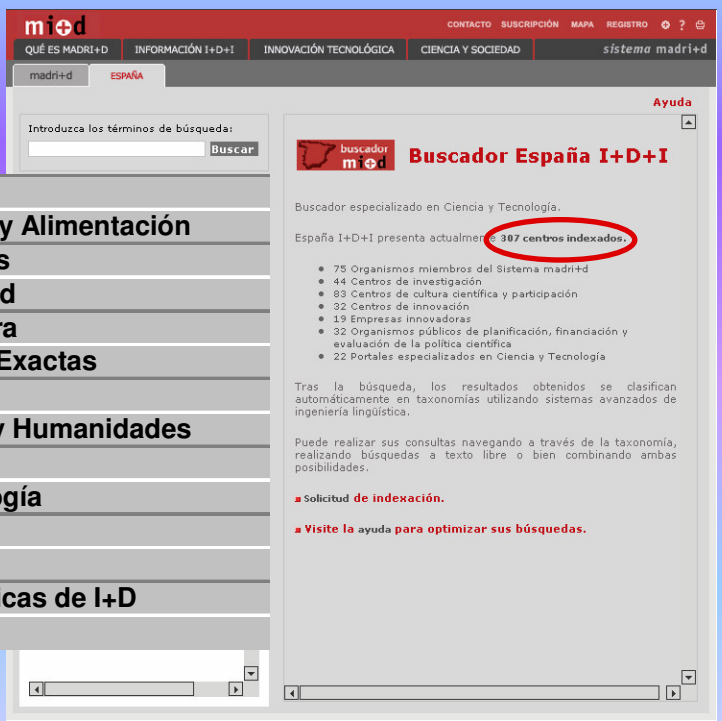
CONTRIBUCIÓN ACADEMICA (Google, miles)			
FORMATOS	GLOBAL	UNIVERSIDADES	
ACROBAT PDF	608.000	114.849	18,9%
MS WORD DOC	119.000	16.898	14,2%
MS POWERPOINT PPT	22.900	4.447	19,4%
POSTSCRIPT PS	13.600	4.193	30,8%

(Datos propios, Julio 2006)

Otros modelos en la Web visible

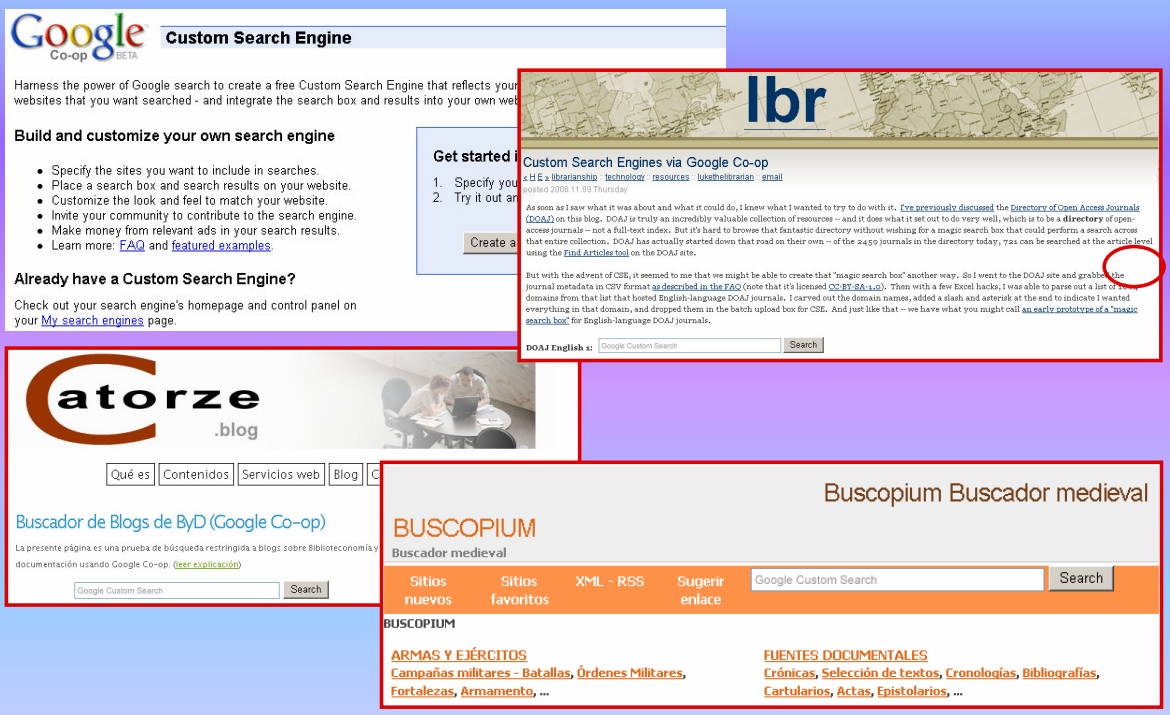
Webpags	España I+D+I
8.156	Agricultura, Pesca y Alimentación
17.995	Ciencias Biológicas
54.798	Ciencias de la Salud
16.275	Ciencias de la Tierra
35.762	Ciencias Físicas y Exactas
13.070	Ciencias Químicas
23.897	Ciencias Sociales y Humanidades
28.781	Energía
49.084	Industria y Tecnología
254	Medidas y Normas
19.976	Medio Ambiente
8.854	Políticas Tecnológicas de I+D
556.455	TOTAL

Noviembre 2006



http://buscador.madrimasd.org/BuscadorMadrimasd/default.asp

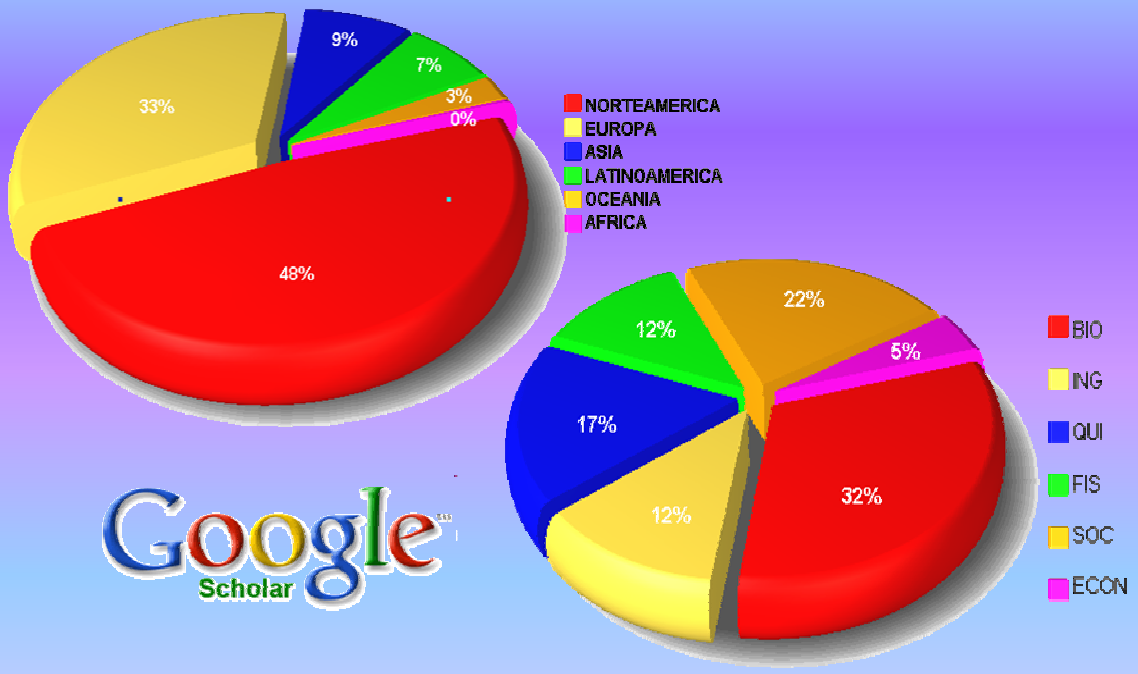
¡No es tan difícil!



Análisis cuantitativo (II)

- Web invisible y privada
  - Referencias bibliográficas:
    - Intermediarios
      - Google Scholar
      - MSN Live Academic
      - Elsevier Scirus
      - ISI Web Citation Index
      - Scopus

Google Académico

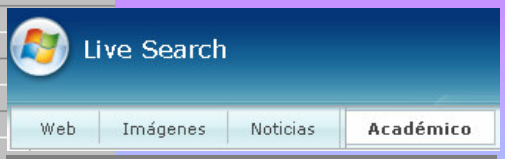


Objetos en dominios universitarios (Datos propios, Julio 2006)

11

Live Academic

LIVE	SCHOLAR	DOMINIOS	
69,72%	39,00%	com+org+net+info...	Editoriales, Asociaciones
14,06%	6,37%	edu	Universidades
8,89%	0,89%	uk	Editoriales/Universidades
1,40%	2,29%	de	
0,63%	7,17%	fr	Universidades
0,51%	0,42%	nl	Editoriales
0,46%	18,92%	us+gov+mil	Informes
0,44%	0,77%	ca	
0,41%	0,33%	ch	
0,36%	0,63%	au	
0,29%	0,47%	it	
0,28%	1,12%	jp	
0,28%	0,33%	se	
0,23%	0,52%	es	
0,19%	0,27%	be	
0,18%	0,27%	at	
0,16%	0,28%	dk	
0,16%	0,29%	fi	
0,14%	0,18%	il	
0,10%	0,16%	gr	
0,09%	0,30%	kr	
0,09%	0,14%	ie	
0,09%	1,93%	br	
0,09%	0,26%	pt	
0,08%	0,28%	cz	
2.739.120	8.245.872	total de ficheros	

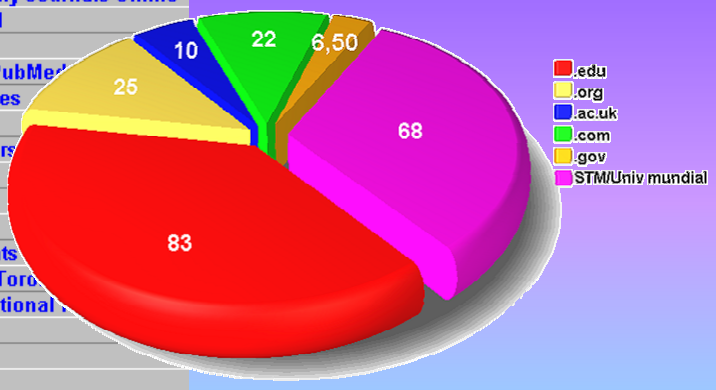


Editoriales	118
Revistas	4.371
Actas de Congresos	2.086

Octubre 2006

Scirus

13.000.000	artículos	BioMed Central
6.600.000	artículos	ScienceDirect
368.000	artículos	PubMed Central
330.000	artículos	Scitation
212.000	artículos	Institute of Physics Publishing
65.500	artículos	Crystallography Journals Online
35.500	artículos	Project Euclid
8.800	artículos	SIAM
17.300.000	referencias	Medline via PubMed
390.000	documentos	Digital Archives
180.000	documentos	RePEc
63.000	documentos	MIT OpenCours
11.000	documentos	DiVa
6.500	documentos	WaY
6.000	documentos	Caltech Coda
5.200	documentos	Organic Eprints
4.400	documentos	University of Toro
2.200	documentos	HKUST Institutional
1.700	documentos	CURATOR
600	documentos	PsyDok
363.500	e-prints	ArXiv.org
2.600	e-prints	Cogprints
13.000.000	patentes	LexisNexis
12.000	informes técnicos	NASA
237.500	tesis	NDLTD



Julio 2006

13

Web Citation Index

378.523	e-prints	90,45%	ARXIV ORG E PRINT ARCHIVE
7.999	e-prints	1,91%	DSPACE MIT
4.305	e-prints/informes	1,03%	CALTECH
2.995	e-prints	0,72%	AUSTRALIAN NATIONAL UNIVERSITY
2.805	e-prints	0,67%	BIOLINE INTERNATIONAL
2.686	e-prints	0,64%	E LIS
2.640	informes	0,63%	DARTMOUTH COLLEGE COMPUTER SCIENCE
2.544	e-prints	0,61%	EPRINTSUQ
1.843	tesis	0,44%	UNIVERSITY OF HELSINKI
1.675	e-prints	0,40%	COGPRINTS COGNITIVE SCIENCES
1.332	informes	0,32%	BUREAU OF LABOR STATISTICS
1.287	informes	0,31%	OECD
1.207	informes	0,29%	COMPUTER LABORATORY UNIVERSITY OF CAMBRIDGE
942	e-prints	0,23%	SCHOLARLYCOMMONS PENN
869	e-prints	0,21%	UNIVERSITY OF TORONTO
641	e-prints	0,15%	QUEPRINTS CRANFIELD UNIVERSITY
598	informes	0,14%	COMPUTER SCIENCE IOWA STATE UNIVERSITY
556	e-prints	0,13%	EUROPEAN RESEARCH
539	e-prints	0,13%	EDINBURGH RESEARCH
439	e-prints	0,10%	DSPACE DREXEL UNIVERSITY
407	e-prints	0,10%	ESPACE CURTIN
345	e-prints	0,08%	DSPACE UNIVERSITY OF ROCHESTER
322	e-prints	0,08%	GLASGOW
209	e-prints	0,05%	DSPACE UNIVERSITY O
202	e-prints	0,05%	UNIVERSITY OF NOTTIN
194	informes	0,05%	EUROPEAN CENTRE FO
177	e-prints	0,04%	MATHEMATICAL INSTIT
143	e-prints	0,03%	OXFORD EPRINTS
54	e-prints	0,01%	DIGITALCOMMONS UTE
27	e-prints	0,01%	HISTORY THEORY OF PSYCHOLOGY
5	e-prints	0,00%	INDIAN INSTITUTE OF SCIENCE RESEARCH PUBLICATIONS

Servicio gratuito proporcionado por la FECYT y el MEC

ISI Web of Knowledge℠

Web Citation Index

Octubre 2006

14

Scopus

15.000	Títulos con revisión por pares
500	Revistas Open Access
700	Actas de Congresos
600	Informes comerciales
125	Series de monografías
28.000.000	Resúmenes
245.000.000	Referencias bibliográficas

WebCites

Right hand panel of Abstract & Refs page

Cited By - Web Sources

26 times

Covered web sources: University repositories (e.g. MIT, DiVA, Caltech), theses & dissertations.

MIT – Open Courseware from the Massachusetts Institute of Technology (MIT)
NDLTD – Networked Digital Library of Theses and Dissertations (170,000 documents remaining 30% will be done by the end of 2006)
DiVA – a collection of Institutional repositories from Scandinavian universities
University of Toronto – The institutional repository of University of Toronto
CalTech – The institutional repository from the California Institute of Technology.

Noviembre 2006

Otros Repositorios

- Institucionales
  - Artículos (E-prints Complutense)
  - Tesis, tesinas (TDCAT)
- Temáticos
  - Pre-prints (arXiv, E-LIS)
  - Artículos (DOIS, RePEC)
  - Artículos + Citas (Scielo)
  - Citas + Artículos (CiteSeer)
  - BBDD bibliográficas con enlaces (Dialnet)
- Revistas electrónicas
  - Directorios + Bases de datos (Highwire)

Enlaces “activos” (I)

Query-Free News Search (2003) (Make Corrections) (2 citations)
Monika Henzinger, Bay-Wei Chang, Brian Milch, Sergey Brin

CiteSeer.IST Scientific Literature Digital Library @ University of Zurich Department of Informatics — Home/Search Bookmark Context Related

View or download:
henzinger.com/monik...p707henzinger.ps
berkeley.edu/~milch/papers...www2003.ps
Cached: PS.gz PS PDF Image Update Help

From: henzinger.com/mon...mpublications (more)
(Enter author homepages)

Abstract: Many daily activities present information in the form of a stream of text, and often people can benefit from additional information on the topic discussed. TV broadcast news can be treated as one such stream of text; in this paper we discuss finding news articles on the web that are relevant to news currently being broadcast. (Update)

Cited by: More
The Anatomy of a News Search Engine - Gulli Dipartimento Di (2005) (Correct)
Universit a di Pisa - Gulli, Ferragina (Correct)

Active bibliography (related documents): More All
0.5: Context-Based Similarity Applied to Retrieval of Relevant Cases - Jurisica (1994) (Correct)
0.0: Collaborative Recommender Agents Based on Case-Based Reasoning.. - Montaner (2003) (Correct)
0.0: A System For Automatic Personalized Tracking of.. - Bollacker, Lawrence.. (1999) (Correct)

Similar documents based on text: More All
1.2: Challenges in Web Search Engines - Henzinger, Motwani, Silverstein (2002) (Correct)
0.6: Least Common Ancestors in Trees and Directed Acyclic.. - Bender, Farach-Colton, .. (2001) (Correct)
0.6: When Experts Agree: Using Non-Affiliated Experts to Rank.. - Bharat, Mihaila (2001) (Correct)

BibTeX entry: (Update)

Monika Henzinger, Bay-Wei Chang, Brian Milch, and Sergey Brin. Queryfree news search. In WWW12, pages 1--10, 2003. http://citeseer.ifi.unizh.ch/henzinger03queryfree.html More

```
@misc( henzinger03queryfree,
  author = "M. Henzinger and B. Chang and B. Milch and S. Brin",
  title = "Queryfree news search",
  text = "Monika Henzinger, Bay-Wei Chang, Brian Milch, and Sergey Brin. Queryfree
    news search. In WWW12, pages 1--10, 2003.",
  year = "2003",
  url = "citeseer.ifi.unizh.ch/henzinger03queryfree.html" )
```

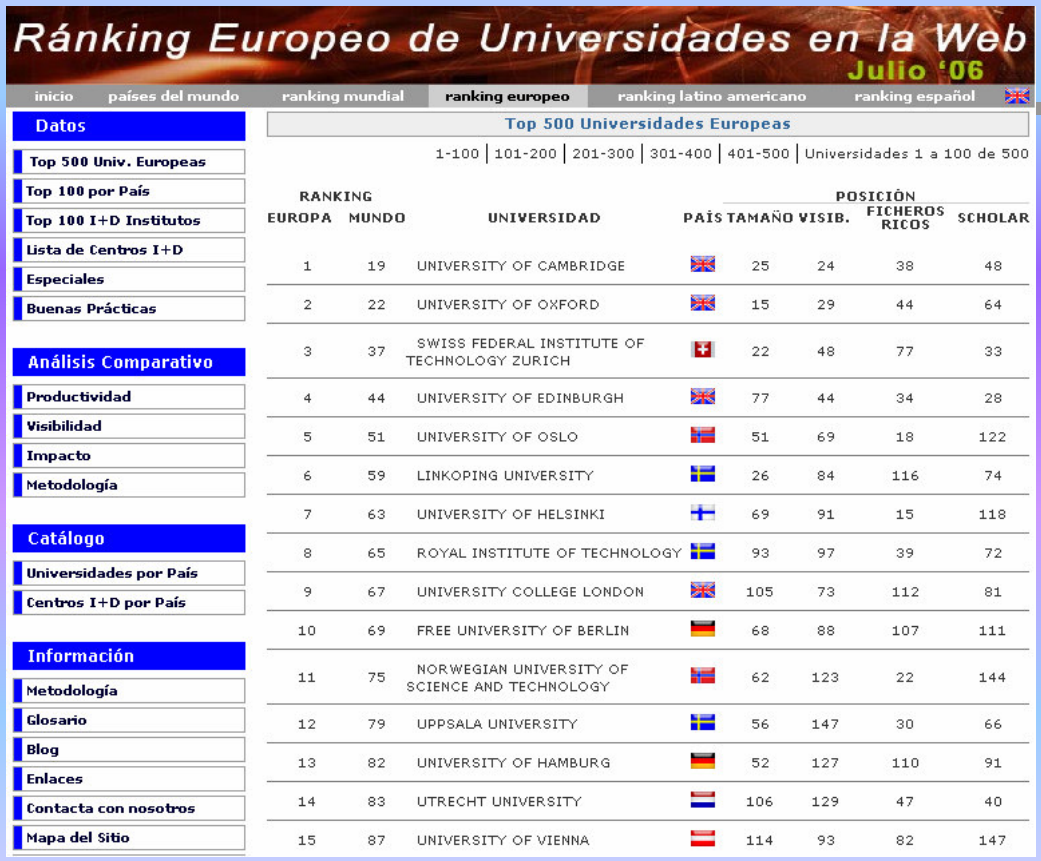
Citations (may not include all citations):
166 Transformation-based error-driven learning and natural langu.. - Brill - 1995
23 What can you do with a web in your pocket - Brin, Motwani et al. - 1998
18 Temporal summaries of news topics - Allan, Gupta et al. - 2001
16 Information access in context (context) - Budzik, Hammond et al. - 2001
15 Domain-specific keyphrase extraction - Frank, Paynter et al. - 1999
11 Learning user information interests through the extraction o.. (context) - Krulwich, Burkey
2 Query-free information retrieval - Hart, Graham - 1997
1 EIA-746-A: Transport of internet uniform resource locator (context) - Alliance - 1998
1 Intercast dying of neglect (context) - Davis - 1997

Documents on the same site (http://www.henzinger.com/monika/mpublications.html): More
Analysis of a Very Large AltaVista Query Log - Silverstein (1998) (Correct)

Enlaces “activos” (II)

Justificación

- Visibilidad
  - Geográfica (países en vías de desarrollo)
  - Socioeconómica (audiencias más amplias)
- Acceso
  - Universal, iniciativas "Open Access"
- Tecnológica
  - Interconexión "profunda"
- Política
  - Herramienta de evaluación
  - Identificando la "brecha digital"



Ránking Europeo de Universidades en la Web

Julio '06

inicio | países del mundo | ranking mundial | ranking europeo | ranking latino americano | ranking español

Datos

- Top 500 Univ. Europeas
- Top 100 por País
- Top 100 I+D Institutos
- Lista de Centros I+D
- Especiales
- Buenas Prácticas

Análisis Comparativo

- Productividad
- Visibilidad
- Impacto
- Metodología

Catálogo

- Universidades por País
- Centros I+D por País

Información

- Metodología
- Glosario
- Blog
- Enlaces
- Contacta con nosotros
- Mapa del Sitio

Top 500 Universidades Europeas

1-100 | 101-200 | 201-300 | 301-400 | 401-500 | Universidades 1 a 100 de 500

RANKING EUROPA	RANKING MUNDO	UNIVERSIDAD	PAÍS	POSICIÓN TAMAÑO	POSICIÓN VISIB.	POSICIÓN FICHEROS RICOS	POSICIÓN SCHOLAR
1	19	UNIVERSITY OF CAMBRIDGE		25	24	38	48
2	22	UNIVERSITY OF OXFORD		15	29	44	64
3	37	SWISS FEDERAL INSTITUTE OF TECHNOLOGY ZURICH		22	48	77	33
4	44	UNIVERSITY OF EDINBURGH		77	44	34	28
5	51	UNIVERSITY OF OSLO		51	69	18	122
6	59	LINKOPING UNIVERSITY		26	84	116	74
7	63	UNIVERSITY OF HELSINKI		69	91	15	118
8	65	ROYAL INSTITUTE OF TECHNOLOGY		93	97	39	72
9	67	UNIVERSITY COLLEGE LONDON		105	73	112	81
10	69	FREE UNIVERSITY OF BERLIN		68	88	107	111
11	75	NORWEGIAN UNIVERSITY OF SCIENCE AND TECHNOLOGY		62	123	22	144
12	79	UPPSALA UNIVERSITY		56	147	30	66
13	82	UNIVERSITY OF HAMBURG		52	127	110	91
14	83	UTRECHT UNIVERSITY		106	129	47	40
15	87	UNIVERSITY OF VIENNA		114	93	82	147

Thinkepi

20

Universidades españolas y CSIC

Thinkepi

RANKING EUROPE	RANKING WORLD	UNIVERSITY	COUNTRY	POSITION SIZE	POSITION VISIB.	POSITION RICH FILES	POSITION SCHOLAR
1	19	UNIVERSITY OF CAMBRIDGE		25	24	38	48
2	22	UNIVERSITY OF OXFORD		15	29	44	64
3	37	SWISS FEDERAL INSTITUTE OF TECHNOLOGY ZURICH		22	48	77	33
42	147	UNIVERSITY COMPLUTENSE MADRID		82	231	253	17
84	248	AUTONOMOUS UNIVERSITY OF BARCELONA		303	315	288	71
86	255	POLYTECHNIC UNIVERSITY OF CATALONIA		255	344	273	139
90	259	POLYTECHNIC UNIVERSITY OF MADRID		278	297	285	297
91	262	UNIVERSITY OF BARCELONA (UB.ES)		336	348	167	120

Ranking Universidades Europeas

Ranking Centros I+D

	INSTITUTE	COUNTRY	POSITION SIZE	POSITION VISIBILITY	POSITION RICH FILES	POSITION SCHOLAR
	NATIONAL INSTITUTES OF HEALTH		3	2	12	1
2	NATIONAL AERONAUTICS AND SPACE ADMINISTRATION		4	4	2	4
8	CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE CNRS		13	15	13	5
8	MAX PLANCK GESELLSCHAFT		9	14	14	16
13	CONSIGLIO NAZIONALE DELLE RICERCHE		17	24	10	32
33	CONSEJO SUPERIOR DE INVESTIGACIONES CIENTIFICAS		54	61	24	27

… ¿el futuro?

- Más contenidos
  - Páginas personales, blogs, wikis, y otros medios informales
  - Integración de las revistas en los repositorios
- Más estructura semántica
  - Valor añadido mediante interconexión hipertextual (super-citas)
  - Web semántica y Web 2.0: Metadatos, etiquetas XML, ontologías RDF
- Menos volatilidad
  - Preservación distribuida

Conclusiones

- Bibliotecas digitales: Escenario rico y diverso
  - Publicación formal: Impacto positivo de las iniciativas *Open Access* (auto-archivo, repositorios institucionales, repositorios temáticos)
  - Publicación informal: Fases de la actividad académica reveladas en la Web
  - Nuevos intermediarios: Acceso universal, nuevos retos para el documentalista
  - Formatos con mayor valor añadido: Web Semántica, interconexión, anotación distribuida, ... Web 2.0, 3.0

¡Gracias!

¿Preguntas?

Más información:

Sede Web	http://internetlab.cindoc.csic.es
Ranking de Universidades	http://www.webometrics.info
Revista-e Cybermetrics	http://www.cindoc.csic.es/cybermetrics