



I Jornadas MAVIR. Madrid, 16-17 Noviembre
2006



Web(o)Metrics

and the Human Language Technologies

Presented by Isidro F. Aguillo & Anselmo Peñas

Red MAVIR (CINDOC-CSIC, UNED)

Agenda

- Definition
- The new mediators for the scholarly communication in the Web
- Quantitative approaches: Web indicators
- Broadening the scenario: Automatic compilation and categorization
- Preliminary results and future developments



Web(o)metrics?

Definition: A quantitative analysis of the usage and presence in the Web by academic institutions, research groups or scientists (introduced by Ingwersen)

Justification: Objective, feasible and comparable methodology for the huge volume of information in the Webspace

Goal: Describe patterns, identify



Why academia?

- **A good candidate:** Strong institutional presence, hierarchical organization, increasing self-archiving practices
- **Practical:** Easy to automate data harvesting, lot of meaningful academic related variables
- **Trend:** Increasing visibility, the next evaluation tool?

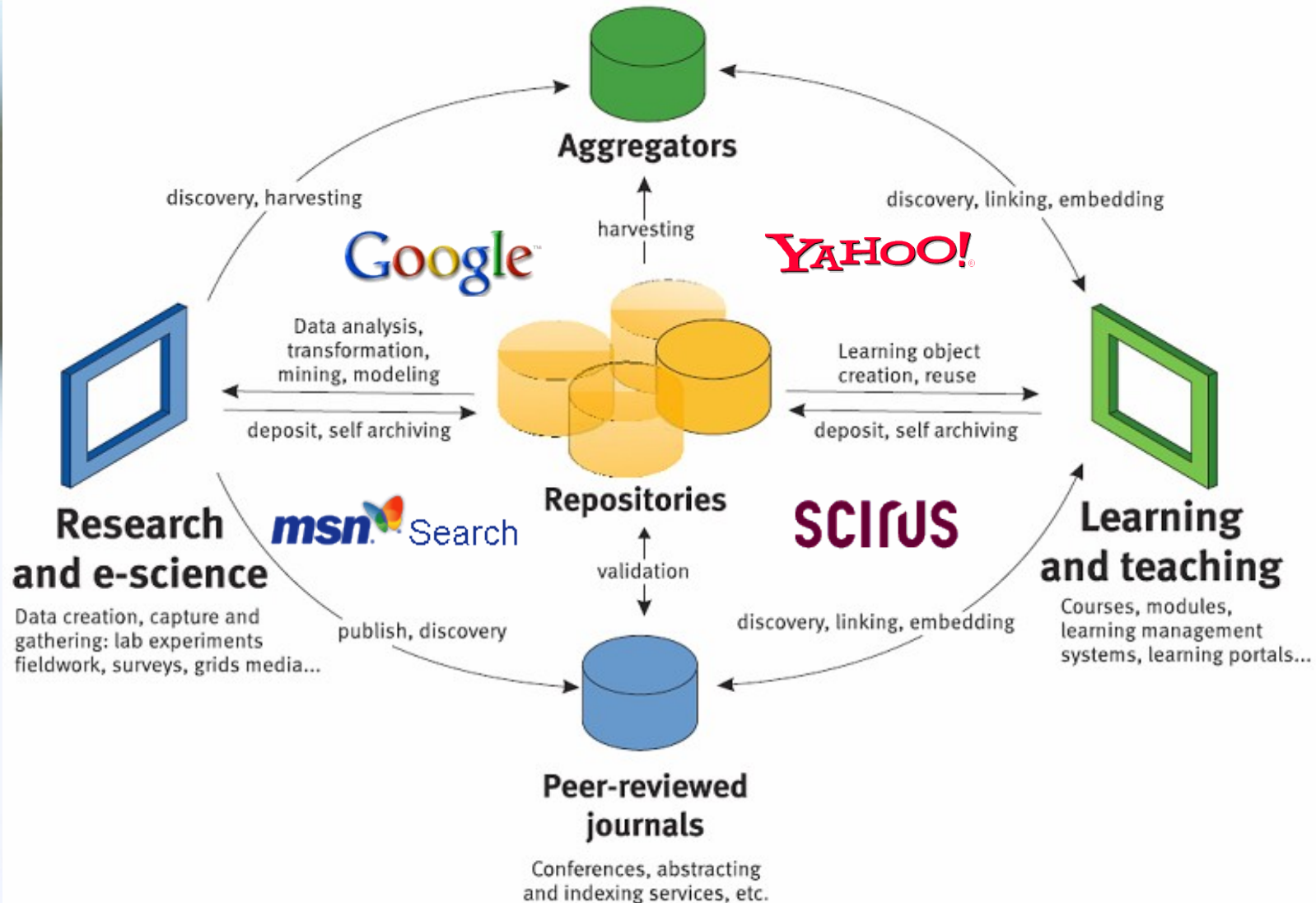


How?

- **Disaggregating:** Academic domains comprise institutes, departments or groups subdomains
- **Measuring:** Search engine crawlers count web objects according to domain filters
- **Explaining:** Link analysis has been proved useful for measuring visibility and impact, correlating with other scientometric indicators



New mediators



Mediators (I): Yahoo

YAHOO! SEARCH

1 - 100 of **about 6,350**

1. [MGH Cancer Center](#)
MGH is a cancer treatment hospital as well as a cancer research and clinical trial institute. ...
Category: [Massachusetts](#) > [Boston](#) > [Massachusetts General Hospital](#)
[cancer.mgh.harvard.edu](#) - 50k - [Cached](#) - [More from this site](#)
2. [Harvard University W.E.B. Du Bois Institute for Afro-American Research](#)
Dedicated to the study of the history, culture, and social institutions of African Americans.
Category: [Massachusetts](#) > [Cambridge](#) > [Harvard University](#) > [Faculty of Arts and Sciences](#)
[web-dubois.fas.harvard.edu](#) - 10k - [Cached](#) - [More from this site](#)
3. [Institute for Theoretical Atomic and Molecular Physics \(ITAMP\)](#)
The Institute is supported by a grant from the National Science Foundation to the Smithsonian Institution and Harvard University.
Category: [Theoretical Physics](#) > [Institutes](#)
[itamp.harvard.edu](#) - 17k - [Cached](#) - [More from this site](#)
4. [Beth Israel Deaconess Cancer Center](#)
Provides cancer information for patients, their families and medical professionals.
Category: [Massachusetts](#) > [Boston](#) > [Hospitals and Medical Centers](#)
[cancercenter.bidmc.harvard.edu](#) - 53k - [Cached](#) - [More from this site](#)
5. [Harvard Medical Web](#)
Category: [Massachusetts](#) > [Cambridge](#) > [Harvard Medical School](#)
[www.med.harvard.edu](#) - 22k - [Cached](#) - [More from this site](#)
6. [Department of Molecular & Cellular Biology](#)
The scientific questions explored in the department span a wide range that include as diverse topics as biochemistry and biophysics, genetics and genomics, development and cell biology, immunology and molecular evolution as well as neurobiology ...
Category: [Massachusetts](#) > [Cambridge](#) > [Harvard University](#) > [Molecular & Cellular Biology](#)
[golgi.harvard.edu](#) - 11k - [Cached](#) - [More from this site](#)



Mediators (II): Live

linkdomain:cancer.mgh.harvard.edu -site:harvard.edu Página 1 de 1.503 resultados

[Online Resources **](#)

SPOHNC. PO Box 53, Locust Valley, NY, 11560-0053, USA. TEL: 1-800-377-0928. FAX: 516-671-8794. info@spohnc.org

www.spohnc.org/resources.html · [Página guardada en caché](#)

[Laboratorios Labin ||| Laboratorio Clínico - Médico](#)

Laboratorios Labin no se hace responsable por el uso de la información que se obtiene en las páginas, tampoco pretende sustituir el consejo ...

www.labinlab.com/enlaces.htm · 30/10/2006 · [Página guardada en caché](#)

[Oncología en Internet](#)

Organización Mundial de Salud. OPS-Organización Panamericana de Salud. Agency for Research on Cancer (IARC)

[cu/links.htm](#) · [Página guardada en caché](#)



Live Search

Web

Imágenes

Noticias

Más ▾

site:cancer.mgh.harvard.edu Página 1 de 115 resultados

discussion group, medical literature search, chemotherapy,

ers medical ... Cancer Links provides direct access to other cancer information services on the Web.

www.meds.com/cancerlinks.html · [Página guardada en caché](#)

[MPRI | Proton Therapy Centers](#)

Your browser does not support script

www.mpri.org/proton_centers.html · [Página guardada en caché](#)

[Mass General Hospital Directory- Find Physicians and Hospital ...](#)

MGH hospital directory provides information on all hospital departments. Use the hospital directory ... Health Departments. MGH specializes in patient care, education and medical research across ...

www.massgeneral.org/departments.html · [Página guardada en caché](#)



Preliminary results

Webometrics Ranking of World Universities July '06

Data

Top 3000 Universities

Premier League

Top USA & Canada

Top Latin America

Top Europe

Top Asia

Top Middle East

Top Oceania

Top Africa

Top 500 R&D Institutes

Research Councils

Distribution by Country

Specials

Best Practices

Comparative Analysis

Productivity

Visibility
















Impact

Methodology

Catalogue

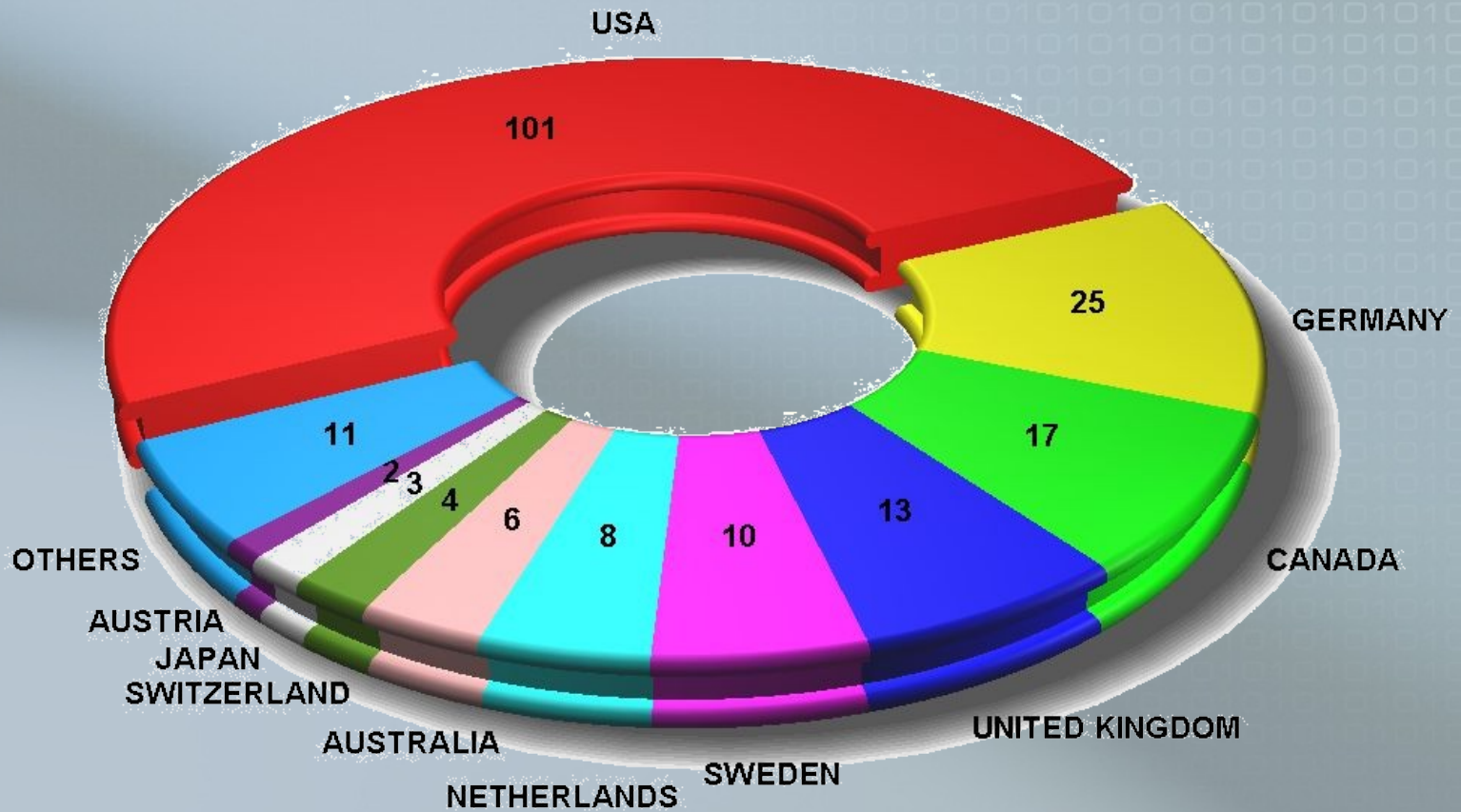
Universities by country

Top 3000 Universities

WORLD RANK	UNIVERSITY	COUNTRY	POSITION			
			SIZE	VISIB.	RICH FILES	SCHOLAR
1	UNIVERSITY OF CALIFORNIA BERKELEY		1	1	2	10
2	HARVARD UNIVERSITY		7	2	5	1
3	MASSACHUSETTS INSTITUTE OF TECHNOLOGY		6	3	3	4
4	STANFORD UNIVERSITY		3	5	1	5
5	UNIVERSITY OF ILLINOIS URBANA CHAMPAIGN		5	7	6	21
6	UNIVERSITY OF MICHIGAN		2	6	13	26
7	CORNELL UNIVERSITY		8	4	9	29
8	UNIVERSITY OF WISCONSIN MADISON		9	9	7	11
9	UNIVERSITY OF TEXAS AUSTIN		12	8	8	19
10	CARNEGIE MELLON UNIVERSITY		4	19	11	12
11	UNIVERSITY OF WASHINGTON		13	12	4	32
12	UNIVERSITY OF CHICAGO		19	10	33	2
13	PENNSYLVANIA STATE UNIVERSITY		10	11	19	45
14	UNIVERSITY OF PENNSYLVANIA		31	13	16	7
15	UNIVERSITY OF MINNESOTA		14	14	10	46



Digital divide !!



Current problems (I)

- **Level of analysis:** University-level web domains with a lot of non-academic information
- **Population size:** Top institutions comprising between $\frac{1}{4}$ and 1 million sub-domains
- **Metadata:** No semantic control of the contents of the main Tags: `<title>`, `<meta>`, `<Hn>`, ...



Current problems (II)

- **Languages:** Webpages available in several local languages: English, German, Spanish, French, Italian and others
- **Geographic allocation:** Groups or departments with different addresses than the top institutions
- **Units comparison:** Subdomains reflect web presence of groups or institutions of different level or nature



Human-Language Technologies

- **Proposal:** Automatic classification of academic sub-domains in 20-25 disciplines
- **Mapping:** Making equivalences between the most extensive, comprehensive and used classifications
- **Learning:** UNESCO codes applied to EICSTES project corpus of EU universities subdomains



Semi-automatic assignment

- **Institutional:** public/private, research universities, institutes, consortia, cooperation
- **Geographic:** Countries, Regions, Developed/Developing, English-/Spanish-speaking
- **Thematic:** Basic/Applied, Biomedicine, Technology, Social Sciences, Humanities, ...



UNESCO codes in EICSTES



European Indicators, Cyberspace and the Science-Technology-Economy System



> Homepage > EU R&D web sites > Finland > University of Helsinki

- ▶ About the project
- ▶ Reports
- ▶ Technical papers
- ▶ Presentations
- ▶ Peer review
- ▶ References
- ▶ Forthcoming events
- ▶ EU R&D web sites
- ▶ Visualization tools
- ▶ Partner area (restricted)
- ▶ Working team
- ▶ Related projects
- ▶ Feedback
- ▶ Disclaimer

Finland



University of Helsinki

<http://www.helsinki.fi/>
 NUTS Code:
 fi161 ::: Uusimaa (Maakunta)
 TOTAL

Sites: 90
 Pages: 95,531
 Links: 33,701
 Objects: 149,077

Accelerator Laboratory

<http://beam.helsinki.fi/>
 Pages: 426 Links: 562 Objects: 2,213

Aerosol and Environmental Physics

<http://www.atm.helsinki.fi/>
 Pages: 10,998 Links: 576 Objects: 16,881

Akka-info

<http://www.helsinki.fi/akka-info/>
 Pages: 205 Links: 103 Objects: 557

Botanical Museum

<http://www.helsinki.fi/kmus/>
 Pages: Links: Objects:

Christina Institute for Women's Studies

http://www.helsinki.fi/~kris_ntk/eindex.html
 Pages: 41 Links: 26 Objects: 54

Cognitive Brain Research Unit

<http://www.cbru.helsinki.fi/>
 Pages: 66 Links: 119 Objects: 115

Computer Science and Engineering

<http://www.cs.helsinki.fi/hecs/e/>
 Pages: 24 Links: 72 Objects: 26

Cultural and Social Anthropology

<http://www.helsinki.fi/hum/antropologia/>
 Pages: 22 Links: 61 Objects: 42

Department of Archaeology

<http://www.helsinki.fi/hum/arla/>
 Pages: 1,381 Links: 137 Objects: 3,325

Department of Art History

<http://www.helsinki.fi/hum/taidehistoria/>
 Pages: 103 Links: 92 Objects: 337



ODP categories

 open directory project

[about dmoz](#) | [suggest URL](#) | [update listing](#) | [become an editor](#)

the entire directory ▾

Top: [Science](#): [Biology](#): [Zoology](#): **Chordates** (179)

- [Herpetology](#) (67)
- [Ichthyology](#) (19)
- [Mammalogy](#) (44)
- [Ornithology](#) (43)

- [Dinosaurs@](#) (82)
- [Tunicates@](#) (4)
- [Vertebrate Paleontology@](#) (92)

See also:

- [Science: Biology: Flora and Fauna: Animalia: Chordata](#) (2,944)

This category in other languages:

[Catalan](#) (31) [French](#) (171) [German](#) (23)
[Italian](#) (77)

- [Cornell University Museum of Vertebrates](#) - Educational and research resource with sections on ichthyology, herpetology, ornithology and mammalogy.
- [Infrared Zoo](#) - Uses infrared photography to show the differences between warm and cold blooded animals.
- [Museum of Vertebrate Zoology](#) - Provides specimen data and archival materials from the collections of the University of California, Berkeley.
- [Museum of Vertebrate Zoology Collections](#) - Searchable database of specimens that includes over 50,000 tissue samples for use in molecular analyses.
- [Society for Northwestern Vertebrate Biology](#) - Devoted to the study of terrestrial vertebrates in the Pacific northwest. Offers publications and reference data.
- [Will's Skull Page](#) - Features images and measurements of mammalian skulls as well as updates and related links.



UDC (Dewey system)



BUBL Information Service

BUBL LINK Catalogue: Selected Internet resources covering all academic subject areas

[Subject Menus](#) | [Countries](#) | [Types](#) | [BUBL UK](#) | [BUBL Archive](#)

Search

[Advanced Search](#)

[A](#) | [B](#) | [C](#) | [D](#) | [E](#) | [F](#) | [G](#) | [H](#) | [I](#) | [J](#) | [K](#) | [L](#) | [M](#) | [N](#) | [O](#) | [P](#) | [Q](#) | [R](#) | [S](#) | [T](#) | [U](#) | [V](#) | [W](#) | [X](#) | [Y](#) | [Z](#)

[000 Generalities](#)

Includes: computing, Internet, libraries, information science

[100 Philosophy and psychology](#)

Includes: ethics, paranormal phenomena

[200 Religion](#)

Includes: bibles, religions of the world

[300 Social sciences](#)

Includes: sociology, politics, economics, law, education

[400 Language](#)

Includes: linguistics, language learning, specific languages

[500 Science and mathematics](#)

Includes: physics, chemistry, earth sciences, biology, zoology

[600 Technology](#)

Includes: medicine, engineering, agriculture, management

[700 The arts](#)

Includes: art, planning, architecture, music, sport

[800 Literature and rhetoric](#)

Includes: literature of specific languages

[900 Geography and history](#)

Includes: travel, genealogy, archaeology

BUBL uses the Dewey Decimal Classification system as the primary organisation structure for its catalogue of Internet resources. The Dewey Decimal Classification is (c) 1996-2005 OCLC Online Computer Library Center. Used with permission.

[E-LIS](#) | [CDLR Projects](#) | [Contacts and Credits](#)

BUBL Information Service, [Centre for Digital Library Research](#), [Strathclyde University](#), Glasgow G1 1XH, Scotland

Tel: 0141 548 4752

Email: bubl@bubl.ac.uk



Criteria and weights

Table I Source of feature and their relative weights

Source of features	Weight
First paragraph	2
Last paragraph	2
Important sentences proposed by Paice (1990)	3
<title>	4
<H1>	2
<H2>	1
	2
	3
<Meta Name = "keywords" or "descriptions">	3
URL	4



Choi, B. & Peng, X. (2004). Dynamic and hierarchical classification of Web pages. **Online Information Review**, 28(2):139-147

Building new indicators

- **In-depth analysis:** Applying HLT for developing more descriptive and finer indicators
 - **link-based**
 - **content-based**
 - **usage-based**
 - **structure-based**
- **Rankings:** Classifying other R&D related organizations



Spanish project proposal

- **QEAVis.** Quantitative Evaluation of Academic Websites Visibility

Research coordinator: María Felisa Verdejo Maillo

- **SUB-PROJECT 1.** Catiex: Multilingual Web Categorization and Information Extraction (UNED)

- **SUB-PROJECT 2.** e-Humanities: Web mediators in the scholarly communication processes (CINDOC-CSIC)



Thank you!

- **Questions?**

- **Contacts:**

Cybermetrics Research Group. CINDOC-
CSIC

<http://internetlab.cindoc.csic.es>

Natural Language Processing and
Information Retrieval Group. UNED

<http://nlp.uned.es/>

