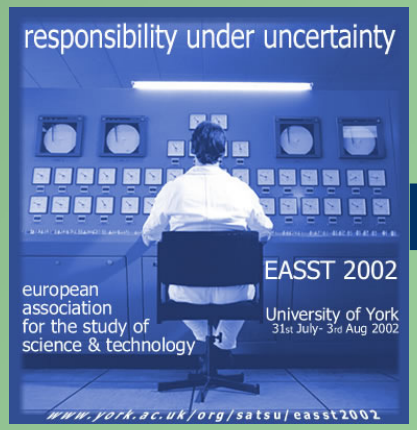
Measuring informal scientific publication in the Web

Isidro F. Aguillo
CINDOC-CSIC
isidro@cindoc.csic.es
SPAIN



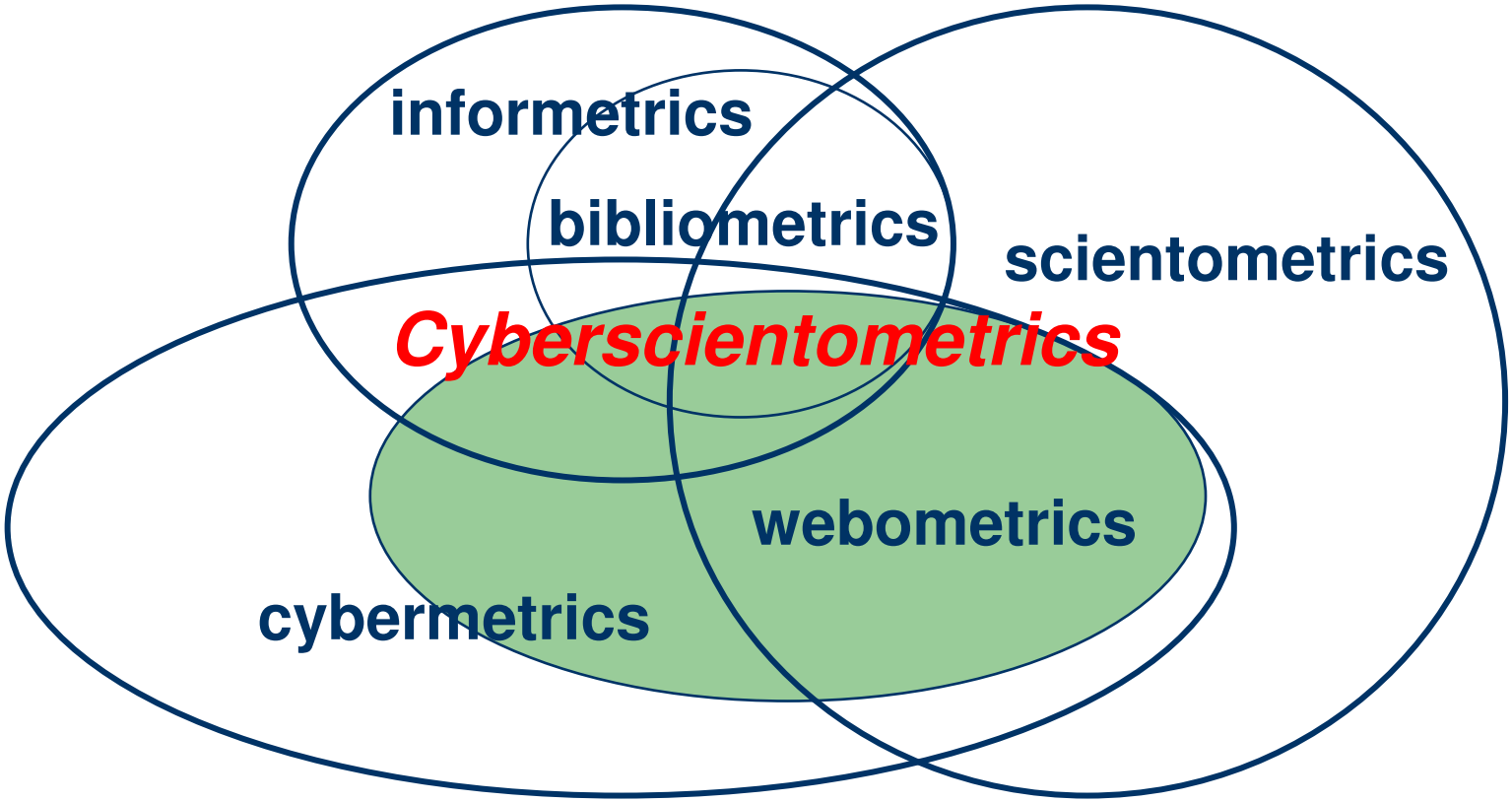
Quantitative approaches

- The presence on the Web of research groups, professors or postgraduate students reflects more activities and results than the traditional formal publication in refereed journals
  - unpublished material, general public contributions, drafts for future papers or book chapters, slides used in conference or seminar presentations, support material for courses or even raw data
- The Web reaches a wider audience than the paper based publications like journals or books.
  - The information published on the Web can be recover by any Internet user worldwide
- The interlinked nature of the Web offer the possibility to discover hidden relationships among different websites
  - Identifying academic communities but showing also economic, industrial, social or cultural relationships

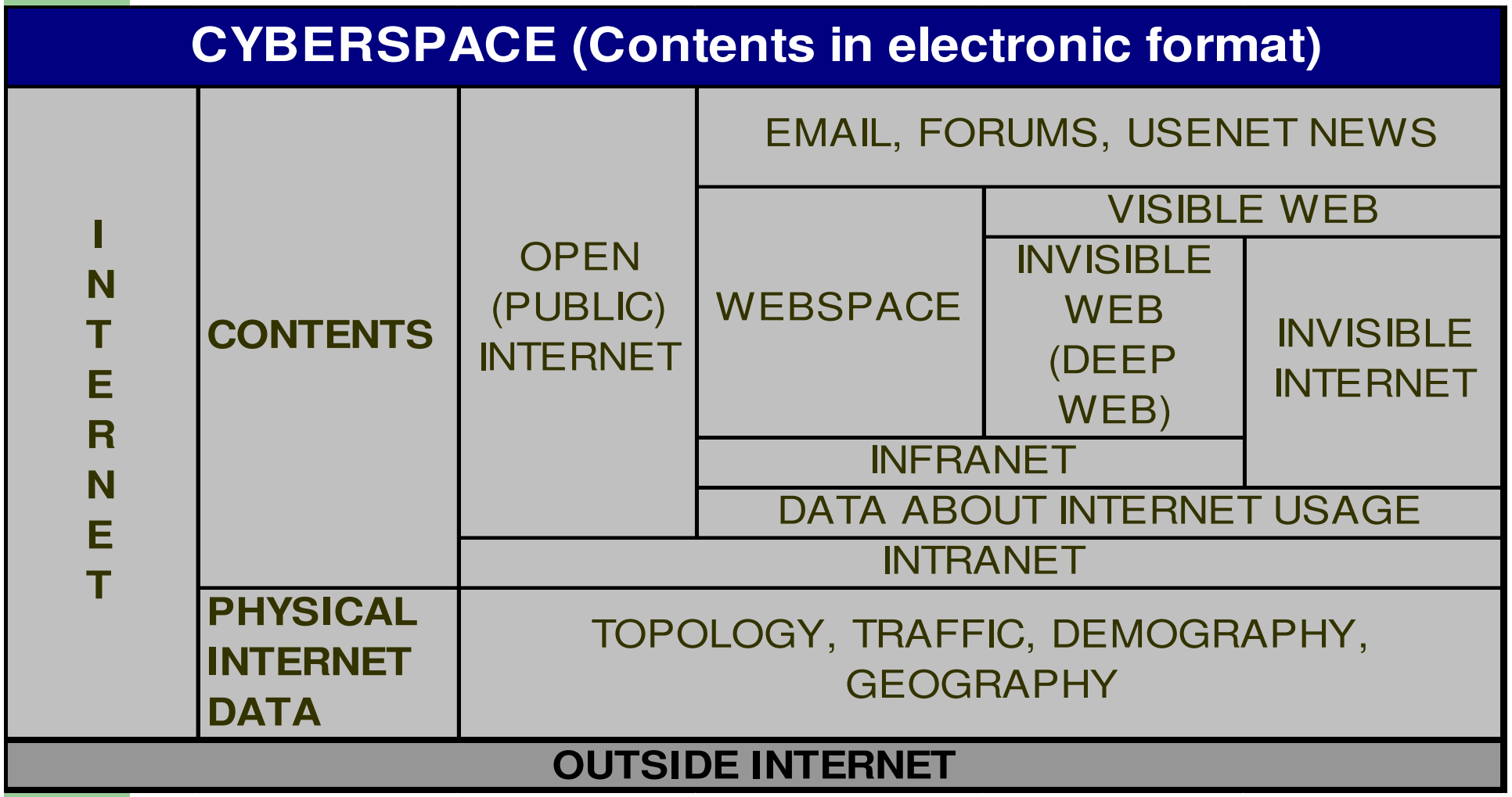
Some definitions

- Cybermetrics is the emerging discipline devoted to the quantitative description of the contents and communication activities that occurs in the cyberspace
- Cyberscientometrics focus on the presence of R&D institutions in the Web and the formal (electronic journals) and informal processes of scholarly communication in the Internet
- Cyberspace=Contents in the Internet

Quantitative disciplines



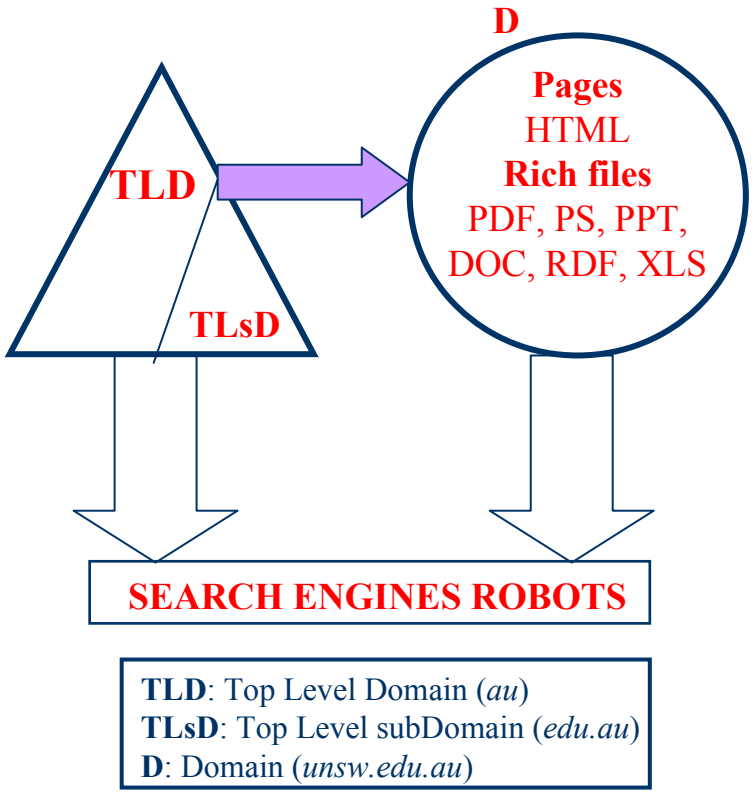
Adapted from L. Björneborn (2002)



INVISIBLE INTERNET			SIZE
INFRANET	Bibliographic Databases	Library catalogues	40,000 webOPACs
		Other bibliographic databases	250,000 databases
	Alphanumeric Databases	Reference: Encyclopaedias, dictionaries	
		Numeric data, statistics	
		Textual, including full text	
INVISIBLE WEB	Orphaned pages		~22%
	Non-textual web pages	Adobe Acrobat, PostScript	300+ millions
		Multimedia files	
	Gateway	Fee or registration required	over 10.000 ejournals
		Documents repositories and electronic journals	
	Active pages	ASP, PHP	500? Millions

2 - 50 times larger than visible web

Methods



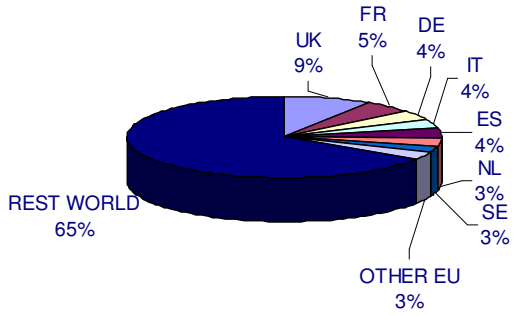
SEARCH ENGINES		
Field delimiters	**FAST**	**Google**
DOMAIN	url.tld	site
SUBDOMAIN	url.host	site
HOST WORD	url.domain	NO
HOSTNAME	url.host	site
URL	url.all	allinurl
SPECIAL	tick pdf	filetype, country (API)

SIZE OF THE WEBSPACE

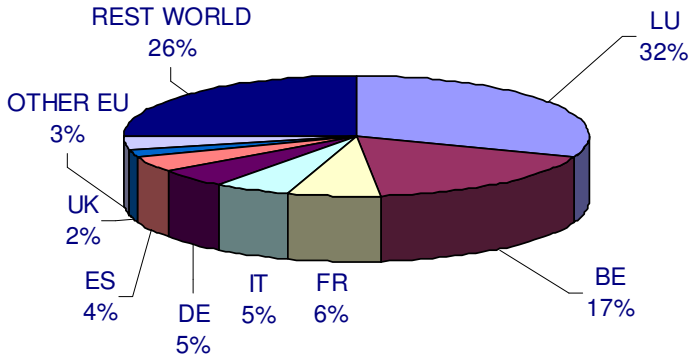
gTLD + US			Europe			Asia-Australasia			America/Africa		
Rank		Webpages	Rank		Webpages	Rank		Webpages	Rank		Webpages
1	com	967.574.482	4	de	107.598.200	5	jp	80.316.887	9	br	32.767.185
2	org	146.541.333	6	uk	62.032.688	10	kr	31.872.332	15	ca	22.173.975
3	net	110.579.260	8	ru	40.508.956	14	au	22.266.917	34	za	4.253.277
7	edu	49.484.142	11	nl	28.234.303	20	cn	13.299.971	35	ar	4.124.638
22	to	12.451.808	12	it	27.995.250	26	tw	10.028.508	40	mx	2.797.374
23	us	12.075.616	13	pl	22.509.107	30	nz	6.269.705	48	cl	1.745.437
25	gov	11.355.141	16	ch	18.042.328	42	il	2.565.176	67	co	679.328
33	nu	4.439.622	17	cz	17.730.451	43	tr	2.490.870	73	pe	419.551
44	cc	2.200.656	18	fr	17.539.647	46	hk	2.167.075	74	ve	410.632
50	mil	1.658.373	19	dk	14.957.171	49	sg	1.699.074	78	uy	336.284
53	vu	1.463.476	21	se	12.700.865	51	my	1.568.214	84	cr	239.202
54	tv	1.386.958	24	at	11.361.273	56	th	1.323.563	90	cu	147.007
55	info	1.363.623	27	no	8.471.288	65	id	749.371	94	ma	132.103
62	ws	895.649	28	fi	7.244.978	69	in	564.260	97	ec	121.433
66	int	693.996	29	es	6.346.719	70	ph	548.936	100	eg	111.090

Source: FAST (July 2002)

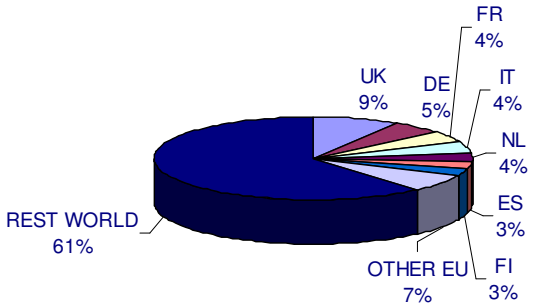
TLD .com



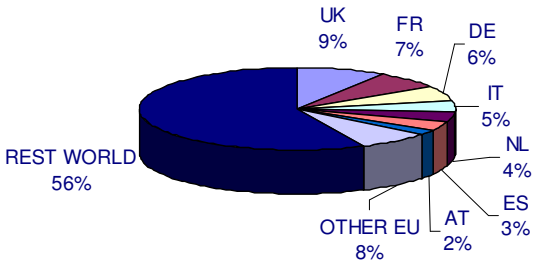
TLD .int



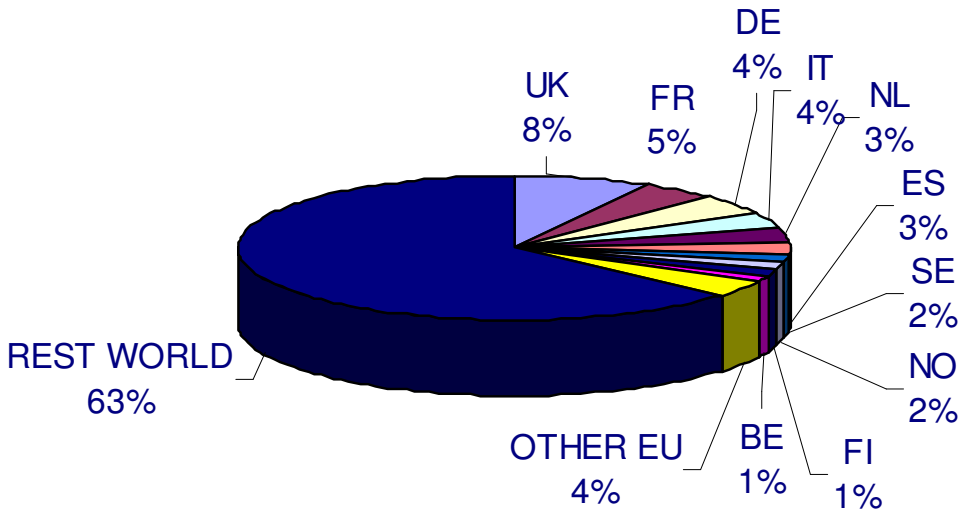
TLD .net



TLD .org



Contribution of EU gTLD



***Source**: API Google, July 2002*

CONTRIBUTION OF INTERNATIONAL DOMAINS TO EU WEBSPACE

COUNTRIES	gTLD						IP number	cTLD
	com	org	net	int	info	edu		
GERMANY	10,1%	6,0%	24,7%	0,3%	0,8%	0,1%	6,5%	68,9%
DENMARK	16,6%	7,2%	28,5%	0,2%	0,1%	0,0%	8,4%	63,0%
GREECE	9,6%	2,3%	14,2%	0,0%	0,0%	0,0%	23,3%	62,5%
AUSTRIA	15,2%	11,3%	34,1%	0,0%	0,6%	0,6%	7,1%	58,8%
PORTUGAL	16,8%	5,8%	29,9%	0,0%	0,1%	0,0%	15,6%	54,5%
NORWAY	20,9%	8,3%	42,5%	0,4%	0,2%	0,0%	5,8%	51,6%
ITALY	20,2%	11,1%	43,6%	0,6%	0,3%	0,1%	6,6%	49,9%
SWEDEN	25,1%	6,8%	40,8%	0,0%	0,1%	0,0%	9,4%	49,7%
EU+NO	20,7%	9,1%	41,2%	0,8%	0,5%	0,1%	9,4%	49,4%

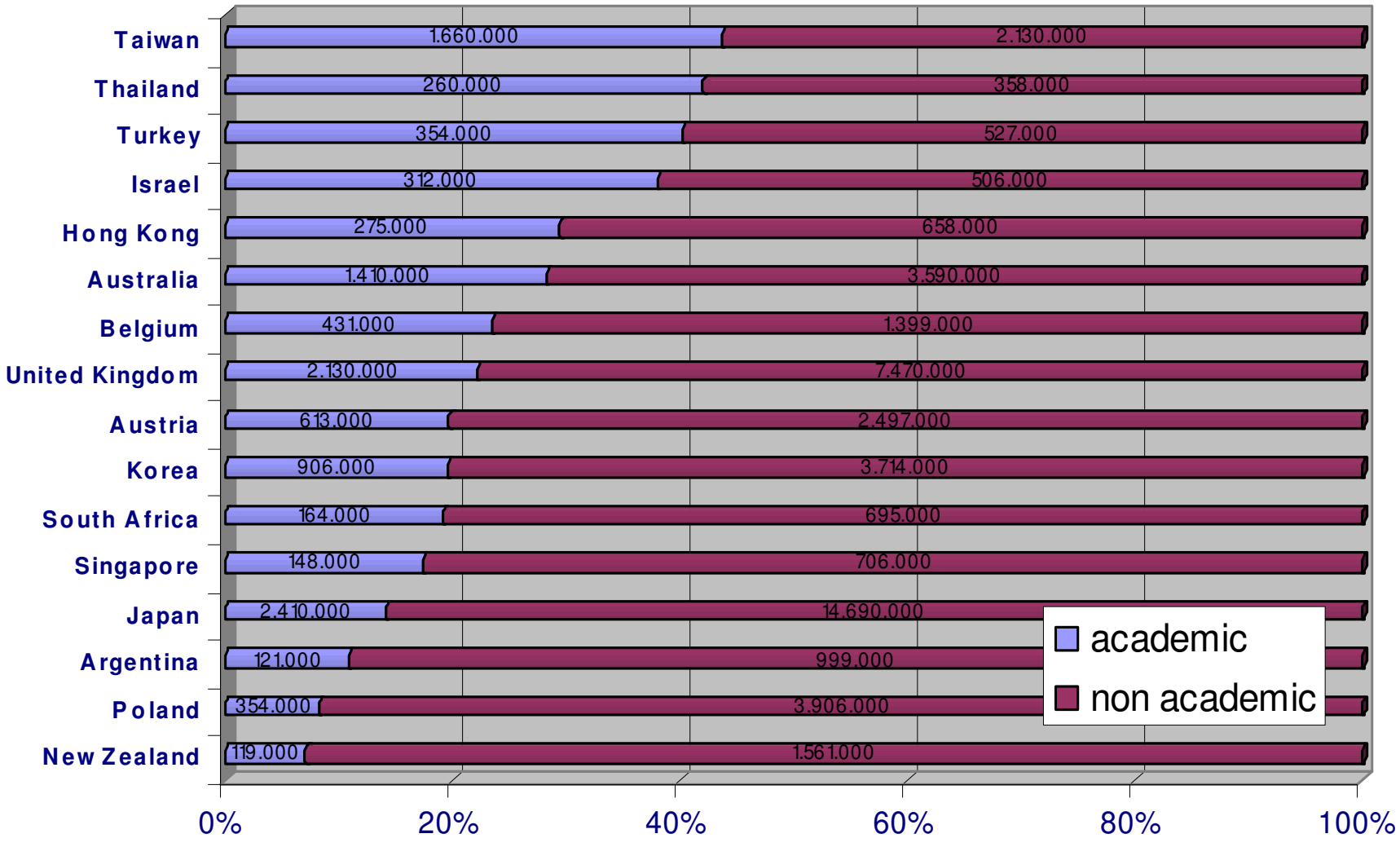
COUNTRIES	gTLD						IP number	cTLD
	com	org	net	int	info	edu		
NETHERLANDS	19,4%	9,6%	40,2%	10,6%	0,1%	0,3%	0,1%	12,0%	47,7%
FINLAND	24,0%	4,9%	48,5%	19,4%	0,0%	0,1%	0,1%	6,1%	45,4%
UNITED KINGDOM	26,8%	11,1%	52,1%	12,9%	0,1%	0,9%	0,0%	9,8%	38,1%
BELGIUM	27,0%	10,0%	49,4%	7,0%	5,4%	0,0%	0,0%	13,4%	37,2%
FRANCE	23,8%	13,6%	49,3%	10,6%	0,6%	0,5%	0,2%	13,8%	36,9%
IRELAND	39,5%	1,5%	42,6%	1,7%	0,0%	0,0%	0,0%	21,8%	35,6%
SPAIN	33,0%	11,9%	59,8%	12,9%	0,8%	0,4%	0,7%	5,1%	35,1%
LUXEMBOURG	2,6%	1,1%	50,6%	0,4%	46,5%	0,0%	0,0%	33,8%	15,6%
WORLD	20,9%	8,1%	43,6%	9,7%	0,4%	0,4%	2,3%	11,6%	44,8%

***Source**: API Google, July 2002*

Informal scholarly communication

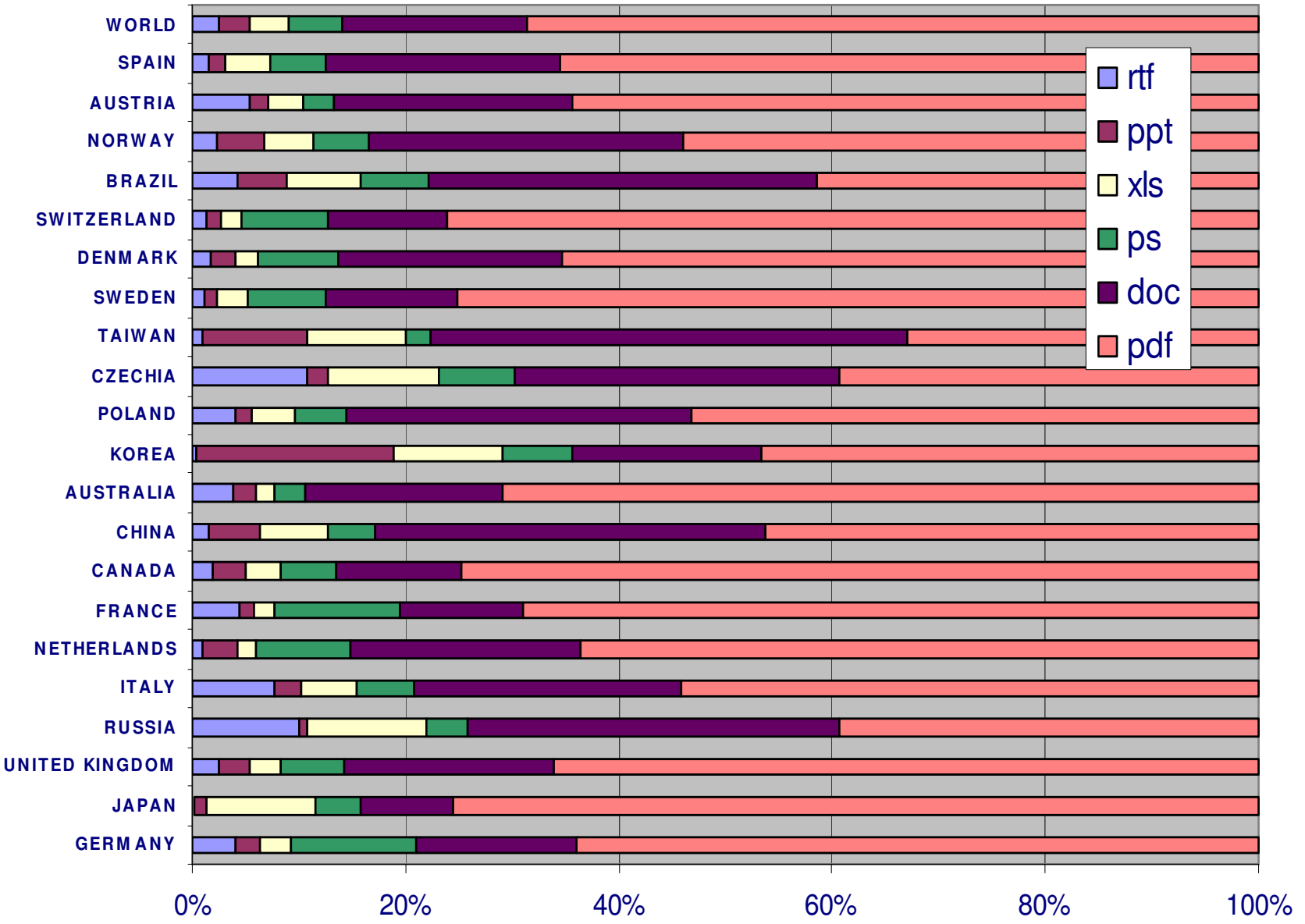
- R&D web size can be estimated from the academic subdomains contribution to webspace
  - Excluding administrative and other non-relevant pages and adding R&D sites under other domains the size of R&D could be over 10%
- Search engines index now rich files (pdf, ps, ppt, doc, xls, rtf) usually associated to informal communication activities in the academic arena
  - Material for students
  - Researchers' personal home pages
    - Papers, conference presentations, drafts, raw data files
  - Departament document archives
  - Electronic libraries (incl. thesis)
  - Subject repositories

Ratio of academic TLsD



API Google, July 2002

Rich files size in selected cTLD



API Google, July 2002

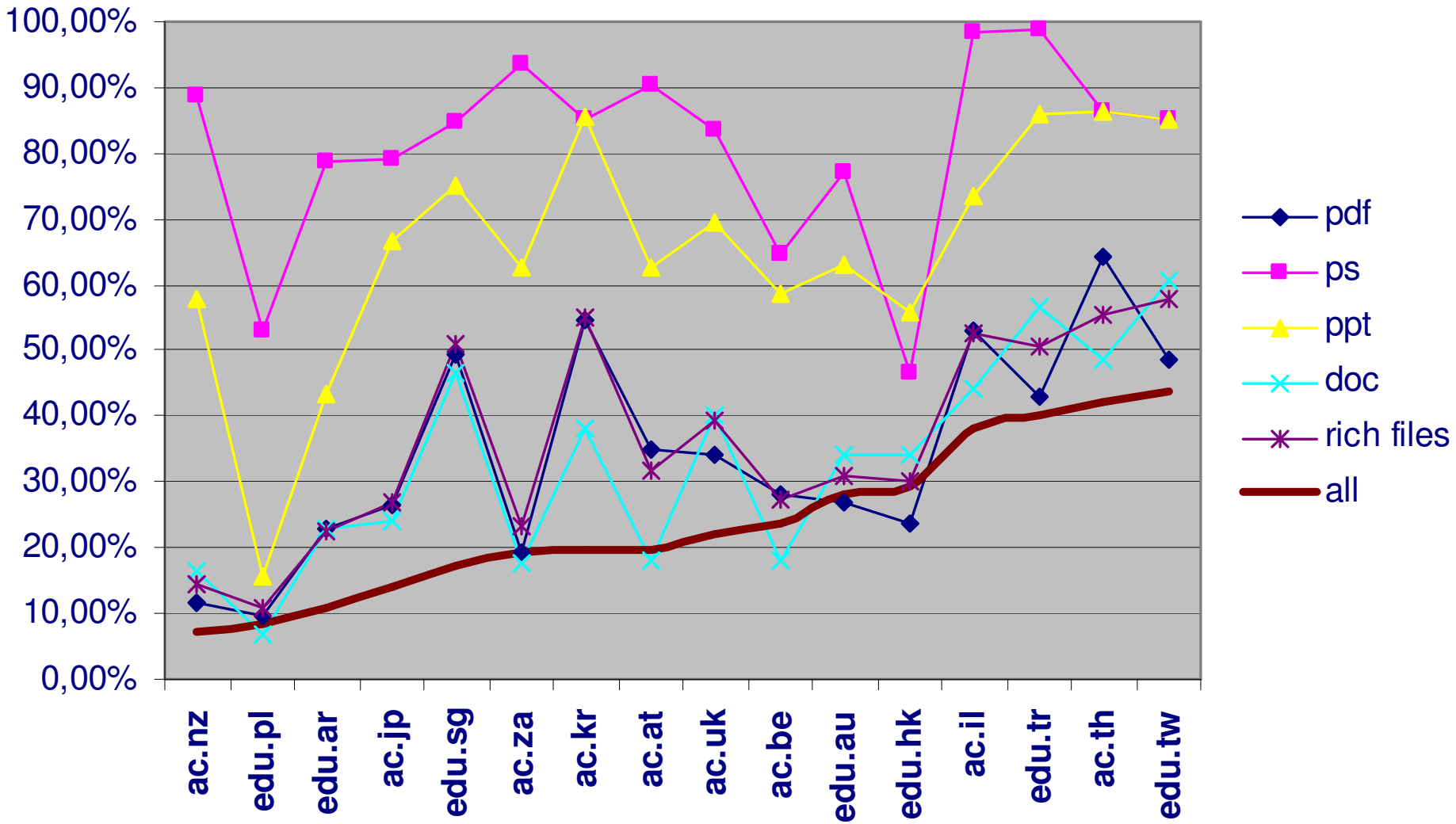
VOLUME OF RICH FILES IN SELECTED cTLD

COUNTRY	rtf	ppt	xls	ps	doc	pdf	non rich
GERMANY	80.900	45.000	57.200	232.000	293.000	1.260.000	18.731.900
JAPAN	2.950	17.100	147.000	59.900	126.000	1.090.000	15.657.050
UNITED KINGDOM	50.900	56.900	61.100	122.000	396.000	1.340.000	7.573.100
RUSSIA	30.700	2.350	34.300	12.000	107.000	121.000	8.822.650
ITALY	67.600	22.900	45.300	48.400	222.000	479.000	5.764.800
NETHERLANDS	5.440	17.600	10.600	49.200	120.000	353.000	5.344.160
FRANCE	35.100	11.200	14.300	94.100	91.600	549.000	4.924.700
CANADA	24.200	41.800	41.300	66.400	153.000	970.000	3.943.300
CHINA	1.850	5.950	7.780	5.490	45.100	57.000	5.076.830
AUSTRALIA	46.600	27.000	21.900	34.200	227.000	871.000	3.772.300
KOREA	955	43.400	24.100	15.300	41.700	110.000	4.384.545
POLAND	9.690	3.450	9.770	11.500	76.900	126.000	4.022.690
CZECHIA	31.600	5.790	30.200	21.200	89.100	115.000	3.747.110
TAIWAN	3.860	36.400	34.500	8.570	168.000	123.000	3.415.670
SWEDEN	8.700	9.630	21.500	56.700	95.500	578.000	2.979.970
DENMARK	6.840	8.900	7.550	28.800	80.200	249.000	3.348.710
SWITZERLAND	10.100	11.800	15.100	63.200	86.700	597.000	2.886.100
BRAZIL	14.800	16.400	24.100	22.400	128.000	145.000	3.239.300
NORWAY	7.850	15.100	16.100	18.000	102.000	186.000	2.824.950
AUSTRIA	21.500	7.220	13.100	12.300	90.400	260.000	2.705.480
SPAIN	7.650	8.300	21.000	25.900	111.000	329.000	2.327.150

API Google, July 2002

Academic TLsD ratio of rich files

API Google, July 2002



LARGEST UNIVERSITIES IN THE WEB

UNIVERSITY	size	pdf	ppt	ps	rtf	doc	xls	rich	
mit.edu	2.130.000	89.800	4.320	52.600	3.430	38.700	4.060	192.910	9,1%
lancs.ac.uk	1.690.000	2.230	428	647	159	1.370	73	4.907	0,3%
ulis.ac.jp	968.000	292	22	132	0	37	0	483	0,0%
harvard.edu	641.000	45.600	974	22.600	801	3.800	571	74.346	11,6%
purdue.edu	555.000	24.500	2.790	5.930	621	7.560	1.020	42.421	7,6%
umb.sk	533.000	33	0	15	12	57	28	145	0,0%
buffalo.edu	488.000	8.800	2.790	1.900	173	2.920	971	17.554	3,6%
mcmaster.ca	449.000	5.960	225	999	65	761	169	8.179	1,8%
stanford.edu	425.000	70.100	3.440	21.500	1.300	5.760	1.570	103.670	24,4%
shu.edu	415.000	634	248	0	0	684	42	1.608	0,4%
cornell.edu	386.000	23.800	2.090	9.250	329	4.400	1.300	41.169	10,7%
ec-lille.fr	379.000	17	13	0	0	12	0	42	0,0%
indiana.edu	321.000	9.500	1.370	5.050	199	3.300	1.470	20.889	6,5%
uibk.ac.at	315.000	4.060	670	212	112	1.060	218	6.332	2,0%
utexas.edu	313.000	38.700	3.190	10.300	595	7.090	3.140	63.015	20,1%
psu.edu	309.000	36.900	6.800	3.610	1.310	13.000	2.390	64.010	20,7%
berkeley.edu	303.000	48.600	5.180	14.600	385	5.740	6.800	81.305	26,8%
sjsu.edu	301.000	5.890	800	67	88	1.850	187	8.882	3,0%
u-tokyo.ac.jp	294.000	16.200	1.210	9.110	108	1.720	394	28.742	9,8%
helsinki.fi	293.000	8.390	821	4.150	637	2.780	1.310	18.088	6,2%
TOTAL (n= 4790)	**85.923.280**	**4.830.531**	**505.888**	**1.151.429**	**175.626**	**1.434.885**	**259.942**	**8.358.301**	**9,7%**

API Google, March 2002

UNIVERSITIES WITH THE HIGHEST RATIO OF RICH FILES (SIZE >50.000)

UNIVERSITY	size	pdf	ppt	ps	rtf	doc	xls	rich	
uni-sb.de	119.000	29.700	492	28.400	33	2.130	20	60.775	51,1%
tamu.edu	189.000	53.000	4.070	4.110	429	7.220	2.800	71.629	37,9%
napier.ac.uk	53.000	1.510	636	128	14.000	2.000	124	18.398	34,7%
iastate.edu	160.000	23.000	2.410	3.020	290	25.000	1.160	54.880	34,3%
lu.se	155.000	49.500	265	1.060	296	1.190	376	52.687	34,0%
ugr.es	52.600	16.700	29	119	51	457	34	17.390	33,1%
cmu.edu	289.000	33.300	3.880	32.300	332	20.200	2.910	92.922	32,2%
rug.ac.be	63.600	15.200	1.250	789	152	2.220	306	19.917	31,3%
uic.edu	289.000	30.600	804	23.900	52	33.500	360	89.216	30,9%
rug.nl	64.400	16.000	514	1.960	137	860	122	19.593	30,4%
ufl.edu	161.000	37.500	2.490	2.470	237	4.590	845	48.132	29,9%
ucm.es	132.000	36.700	73	409	80	1.230	18	38.510	29,2%
lth.se	55.800	9.990	81	4.430	151	1.260	52	15.964	28,6%
washington.edu	210.000	34.400	3.400	11.400	497	7.990	2.000	59.687	28,4%
uark.edu	57.400	13.600	753	57	191	1.100	277	15.978	27,8%
alaska.edu	58.100	12.300	190	1.820	48	1.600	155	16.113	27,7%
vt.edu	226.000	48.300	2.500	1.500	489	7.930	1.680	62.399	27,6%
usp.br	106.000	16.300	1.010	8.460	366	2.550	368	29.054	27,4%
ksu.edu	68.600	11.800	1.750	1.720	228	2.290	846	18.634	27,2%
berkeley.edu	303.000	48.600	5.180	14.600	385	5.740	6.800	81.305	26,8%

API Google, March 2002

RELATIVE POSITION OF UNIVERSITIES ACCORDING TO THE VOLUME OF RICH FILES

USA	RANK							REST OF THE WORLD	RANK						
	PDF	PPT	PS	RTF	DOC	XLS	RICH		PDF	PPT	PS	RTF	DOC	XLS	RICH
mit.edu	1	4	1	3	1	5	**1**	uni-sb.de	19	267	3	792	134	1146	**12**
stanford.edu	2	12	6	7	21	25	**2**	lu.se	4	453	235	114	315	150	**17**
cmu.edu	15	7	2	100	4	9	**3**	ucm.es	13	1022	443	481	308	1195	**25**
uic.edu	18	152	4	635	2	159	**4**	liu.se	29	281	50	365	67	236	**33**
berkeley.edu	5	2	7	84	23	1	**5**	ethz.ch	31	374	24	111	361	184	**35**
harvard.edu	7	108	5	25	49	96	**6**	usp.br	48	103	20	94	96	153	**38**
tamu.edu	3	6	64	73	13	10	**7**	u-tokyo.ac.jp	45	83	17	380	192	143	**39**
psu.edu	12	1	76	6	5	12	**8**	kth.se	39	250	31	299	92	176	**44**
umich.edu	8	23	13	23	7	13	**9**	utoronto.ca	41	70	119	81	105	50	**49**
vt.edu	6	24	188	57	9	23	**10**	chalmers.se	52	413	29	322	230	329	**52**
utexas.edu	11	15	11	46	14	7	**11**	ntnu.no	64	137	43	120	37	94	**53**
umn.edu	9	5	25	44	11	27	**13**	ubc.ca	49	222	67	147	136	232	**58**
washington.edu	14	13	9	56	8	15	**14**	hut.fi	67	158	34	101	135	106	**59**
uiuc.edu	16	3	8	108	29	19	**15**	tu-chemnitz.de	173	118	84	33	6	65	**61**
iastate.edu	27	29	93	117	3	37	**16**	uu.se	86	515	18	160	220	342	**62**
wisc.edu	17	26	10	62	24	22	**18**	cam.ac.uk	102	414	12	177	357	625	**63**
ufl.edu	10	25	117	155	34	62	**19**	tudelft.nl	79	14	169	400	40	125	**65**
purdue.edu	25	19	35	42	10	44	**20**	uni-karlsruhe.de	101	373	14	404	271	463	**67**
arizona.edu	21	34	27	118	33	42	**21**	uni-stuttgart.de	72	331	30	317	243	490	**68**
cornell.edu	26	35	16	103	36	33	**22**	snu.ac.kr	82	8	105	313	132	180	**69**

API Google, March 2002

Conclusions

- Internet offers the possibility to describe in great detail the R&D activity, specially the information not usually published in scientific journals
- Informal communication can require the use of special (rich) file types, currently indexed by major search engines *(Google, Fast)*
- Results obtained from field delimiters show that:
  - Academic subdomains are a relevant part of the Webspace
  - The ratio of rich files is higher in these academic subdomains
  - The volume of rich files can be an indicator of productivity for the universities
  - Several universities are organising large repositories of documents in order to disseminate scientific results

THANK YOU