Creative Commons Statistics from the CC-Monitor Project

Giorgos Cheliotis

School of Information Systems

Singapore Management University

giorgos@smu.edu.sg

Based on a presentation at the iCommons Summit,
Dubrovnik, June 14-17, 2007

iSummit
2007

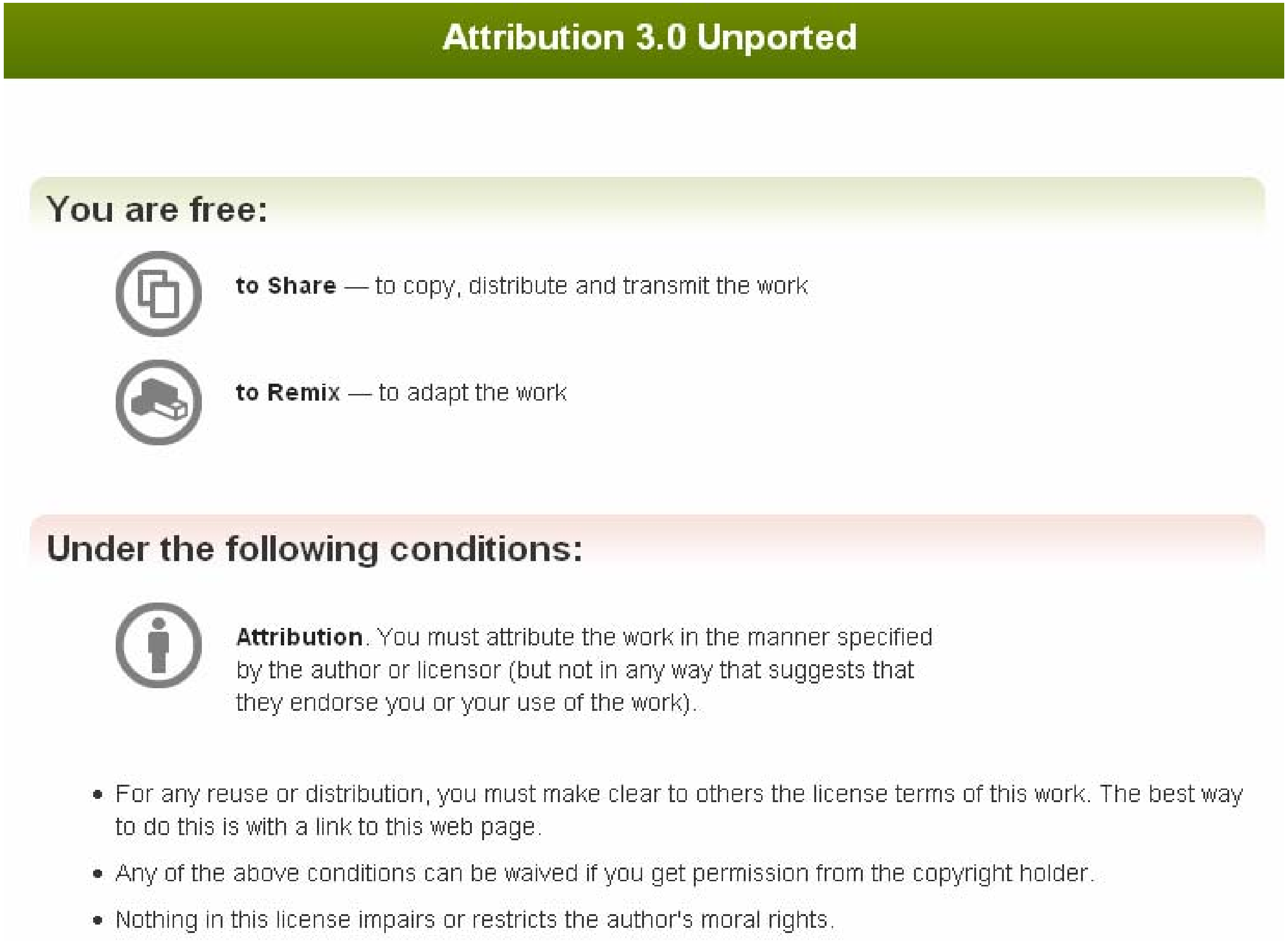
License (1/2)

This material is released early due to high demand and for the benefit of the Creative Commons community – researchers and academics interested in the details of the work are advised to contact giorgos@smu.edu.sg, as the related research is ongoing and currently in the process of being published.

See next page for license details…

Data presented herein was collected in early 2007. It is based on (imprecise) search engine estimates and is therefore only indicative of the real quantities whose size we are attempting to assess.

License (2/2)



Attribution 3.0 Unported

You are free:

- **to Share** — to copy, distribute and transmit the work
- **to Remix** — to adapt the work

Under the following conditions:

Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).

- For any reuse or distribution, you must make clear to others the license terms of this work. The best way to do this is with a link to this web page.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- Nothing in this license impairs or restricts the author's moral rights.

Data presented herein was collected in early 2007. It is based on (imprecise) search engine estimates and is therefore only indicative of the real quantities whose size we are attempting to assess.

Motivation for our study of CC

Before CC most content authors were faced with a binary decision problem: reserve all rights (default copyright protection) or give it all up (public domain)

With CC for the first time we can observe large numbers of users making conscious licensing decisions for their content!

First-level questions

- How many authors use CC?
- Who are they?
- Which licenses do they prefer?
- What is the impact of their choice?
- How do jurisdictions compare?

The really important questions

- How strong is CC adoption?
- How do users value different rights?
- Which factors influence this valuation?
- What are suitable business models for CC content?

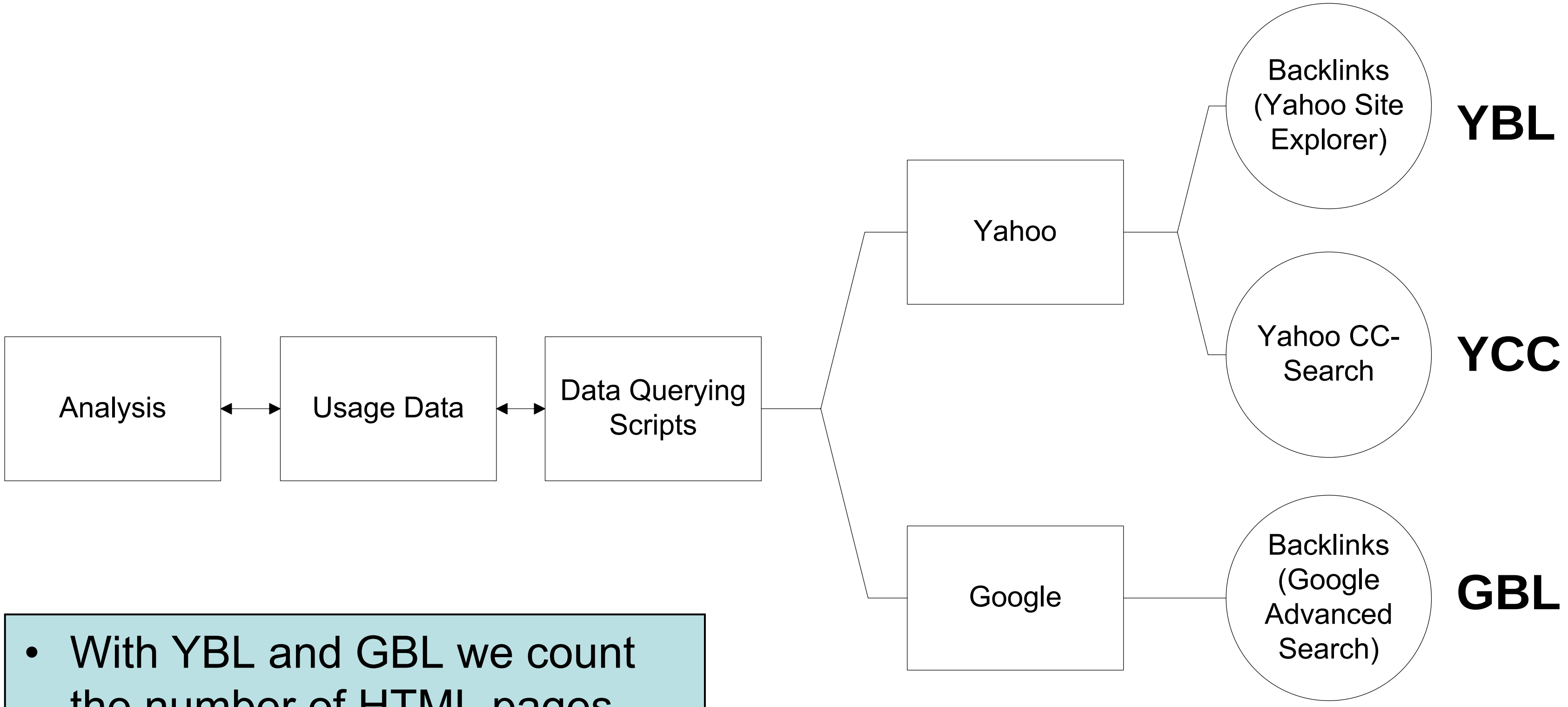
Data presented herein was collected in early 2007. It is based on (imprecise) search engine estimates and is therefore only indicative of the real quantities whose size we are attempting to assess.

Estimates of CC license popularity

- Some data has been made available online by Mike Linksvayer and Christian Ahlert (Openbusiness), in a paper by Zachary Katz, and in a user survey documented in the PhD dissertation of Minjeong Kim
- Most data collection efforts based on Yahoo and Google search results
- Some observations made in the past:
  - Non-BY licenses barely used (and therefore dropped)
  - Total of millions of CC-licensed items (various estimates)
  - NC licenses more popular
  - SA and ND also popular attributes
  - Media type may play a role in licensing (music more liberal)

Data presented herein was collected in early 2007. It is based on (imprecise) search engine estimates and is therefore only indicative of the real quantities whose size we are attempting to assess.

Data collection process (simplified)

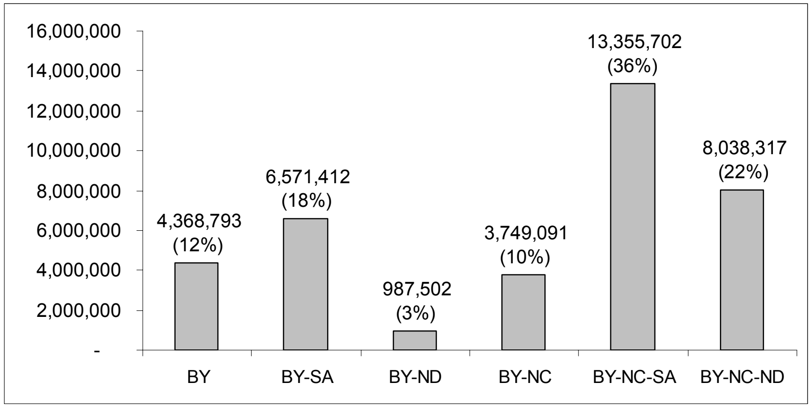


- With YBL and GBL we count the number of HTML pages linking to each CC-Deed page
- With YCC we use Yahoo’s search for CC metadata

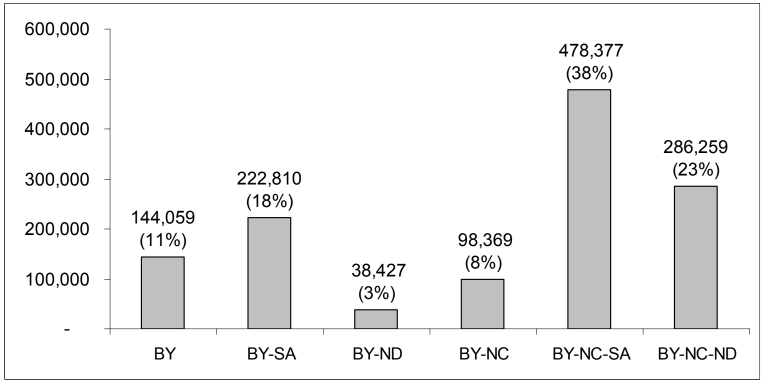
Data presented herein was collected in early 2007. It is based on (imprecise) search engine estimates and is therefore only indicative of the real quantities whose size we are attempting to assess.

Total volume and license mix

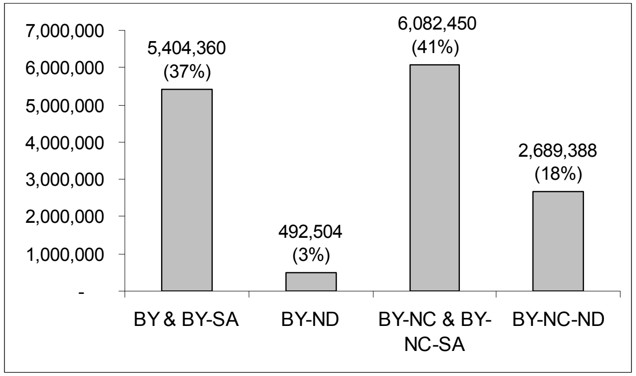
YBL (Total: 37.1m)



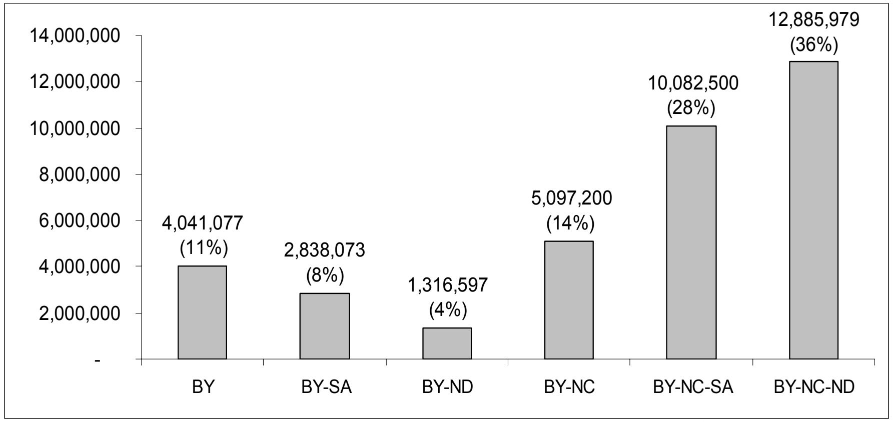
GBL (Total: 1.2m)



YCC (Total: 14.4m)



Flickr (Total: 36.3m)



Data presented herein was collected in early 2007. It is based on (imprecise) search engine estimates and is therefore only indicative of the real quantities whose size we are attempting to assess.

Key observations

- Greatly varying estimates of size of total CC content pool
- However, backlink search with both Yahoo and Google yields an almost identical license mix. In this mix:
  - 70% of the licenses allow non-commercial use only (NC)
  - Share-Alike (SA) also a very popular attribute, present in over 50% of CC-licensed items (though SA is anyhow self-propagating)
  - 25% of the licenses include the ND restriction

- Generally, two groups of content visible, with one group being licensed under clearly more liberal terms and the other under more restrictive terms

- BY-ND unpopular in all measurements, although many items licensed under BY-NC-ND; various interpretations possible

Data presented herein was collected in early 2007. It is based on (imprecise) search engine estimates and is therefore only indicative of the real quantities whose size we are attempting to assess.

Reconciling Flickr and search data

Observations

- Flickr claims to host 36 million CC-licensed items
- According to YBL search results the total CC pool is 37 million items
- Flickr appears to host the bulk of CC content
- Flickr license distribution is U-shaped vs. bimodal distribution of YBL/GBL/YCC (possibly because photographers license differently)

Question

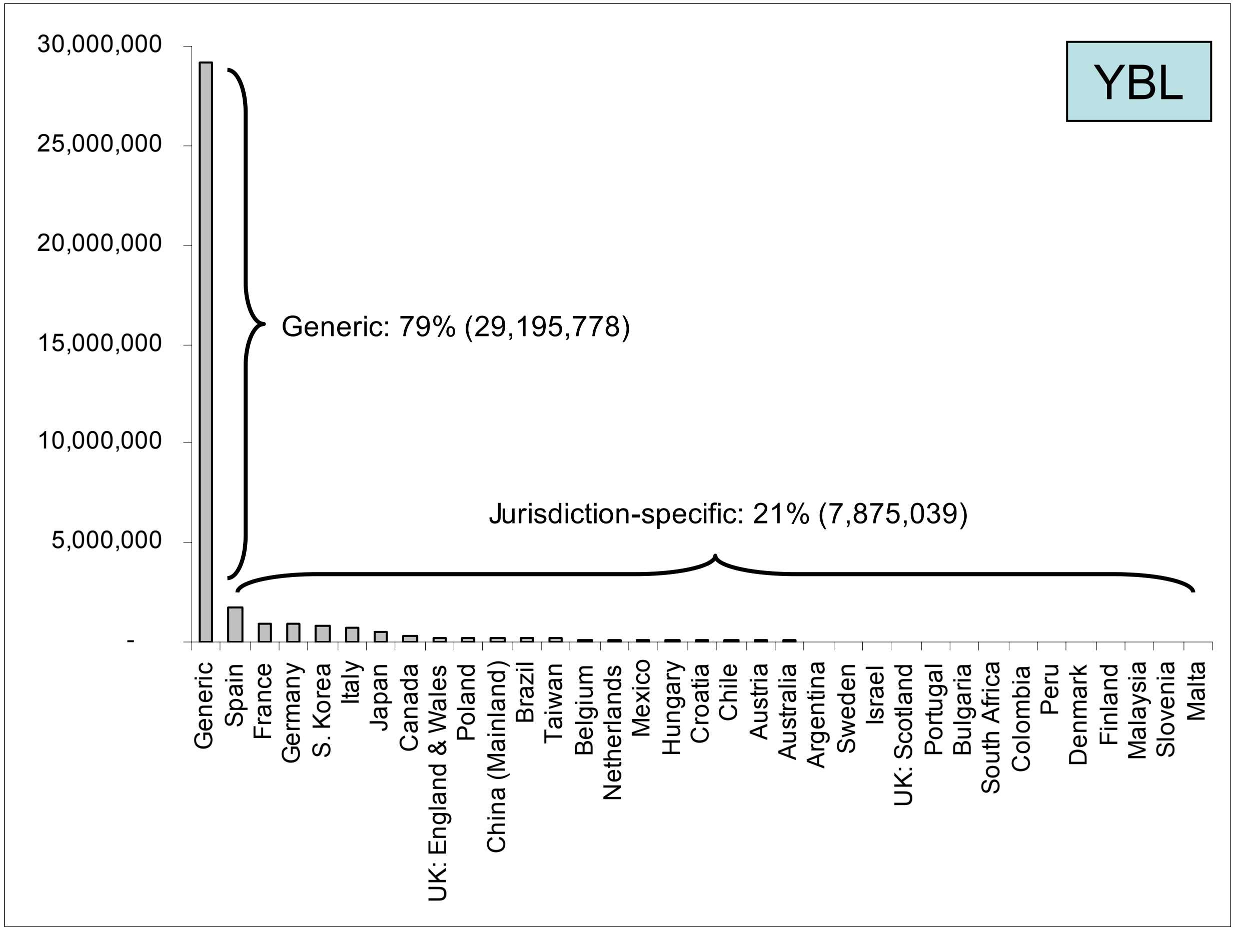
- *How many more CC-licensed items must there be outside Flickr for the Flickr data to be consistent with the search data?*
- The solution to a simple linear optimization problem gives that there must be at least 25,500,000 CC-licensed items outside Flickr!

Grand total: 60+ million CC-licensed items online

Data presented herein was collected in early 2007. It is based on (imprecise) search engine estimates and is therefore only indicative of the real quantities whose size we are attempting to assess.

Volume Generic vs. Jurisdictions

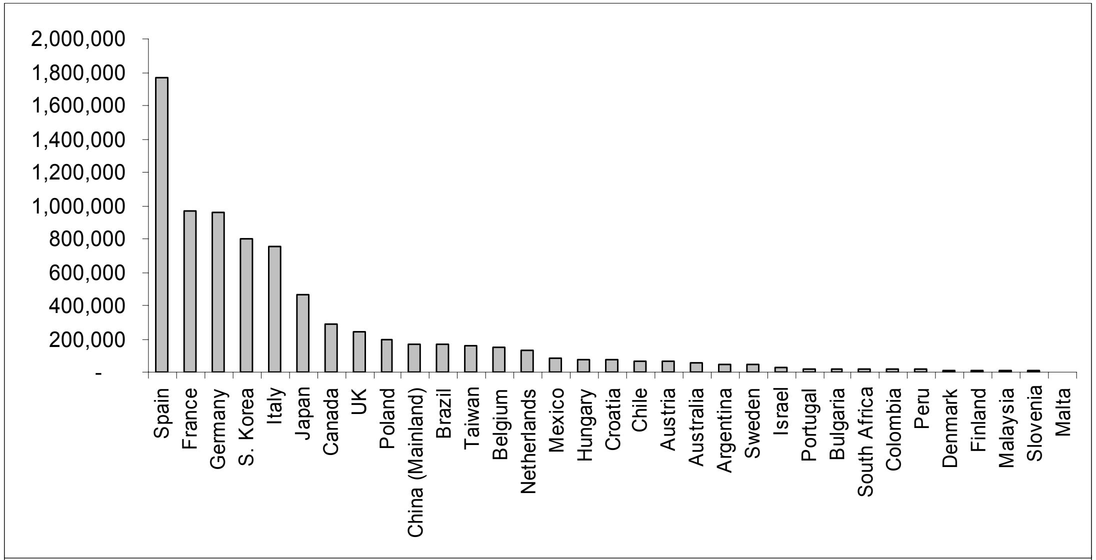
- 80% generic (unported), 20% jurisdiction-specific licenses
- Generic historically the only license
- Jurisdiction-specific expected to grow at least as fast as generic
- “Long tail” is 8 million items, non-negligible



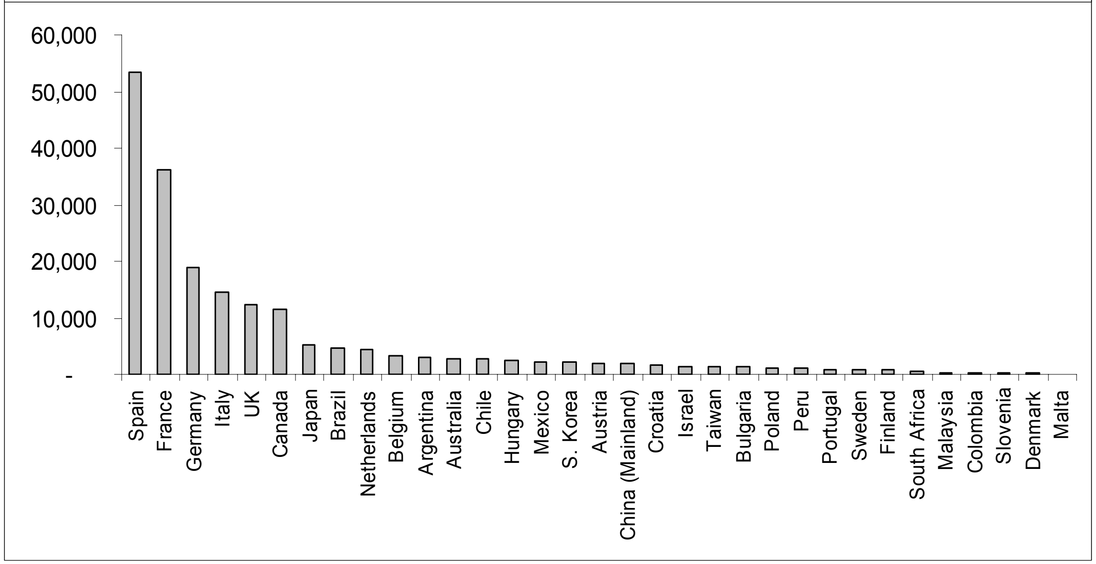
Data presented herein was collected in early 2007. It is based on (imprecise) search engine estimates and is therefore only indicative of the real quantities whose size we are attempting to assess.

Volume per jurisdiction

YBL



GBL



Note: Jurisdiction volume should not be hastily interpreted as country volume since several users may be using the unported licenses or those of another jurisdiction!

Highly correlated

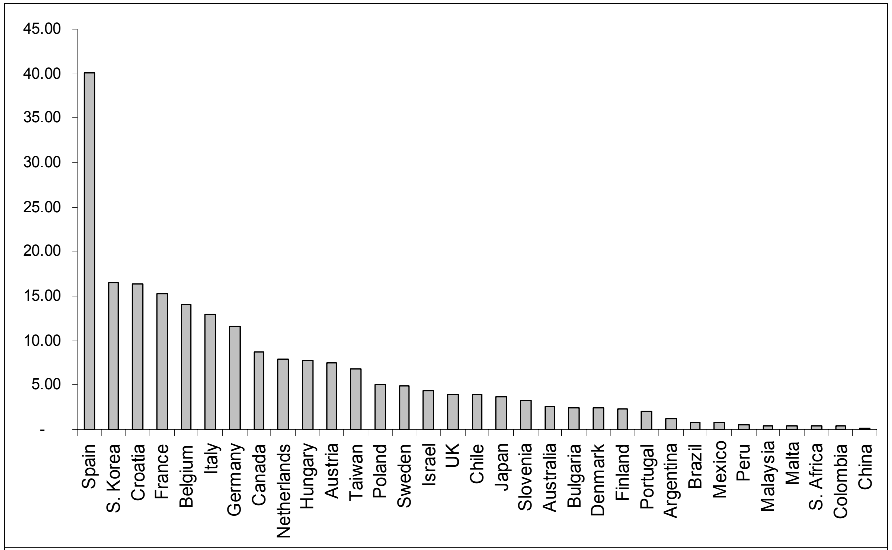
Note: Date of introduction of CC in jurisdiction not taken into account

Note: UK jurisdictions grouped together in this chart

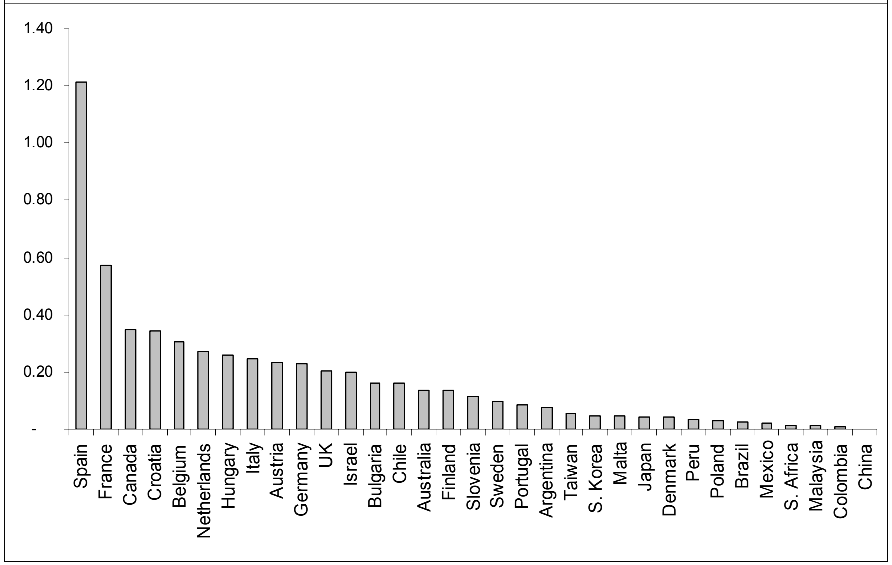
Data presented herein was collected in early 2007. It is based on (imprecise) search engine estimates and is therefore only indicative of the real quantities whose size we are attempting to assess.

Volume per 1000 inhabitants

YBL



GBL



Note: Jurisdiction volume should not be hastily interpreted as country volume since several users may be using the unported licenses or those of another jurisdiction!

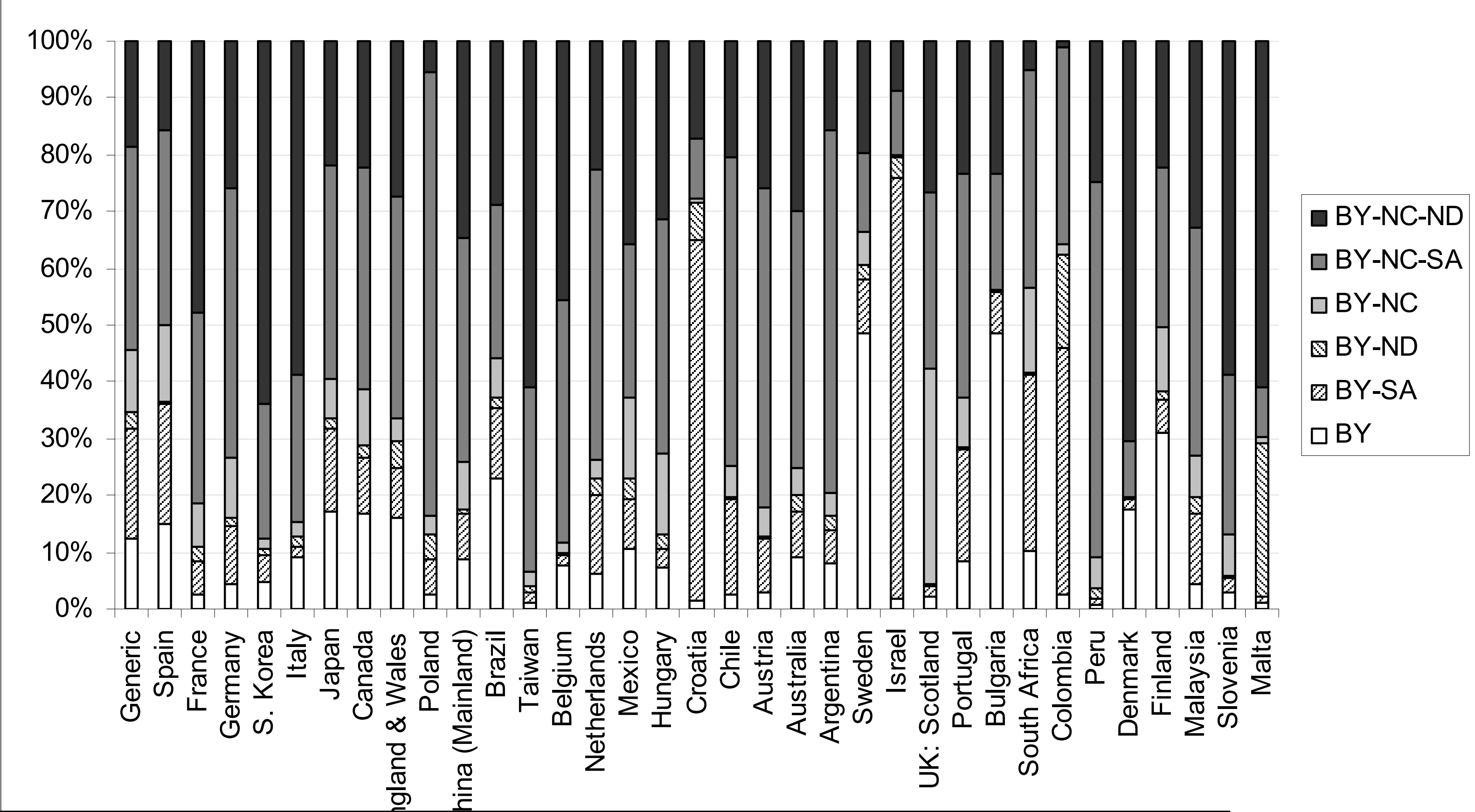
Highly correlated

Note: Date of introduction of CC in jurisdiction not taken into account

Note: UK jurisdictions grouped together in this chart

Data presented herein was collected in early 2007. It is based on (imprecise) search engine estimates and is therefore only indicative of the real quantities whose size we are attempting to assess.

License mix per jurisdiction



- Significant variations, cause unclear
- Careful interpretation needed (jurisdictions ≠ countries, also very different “sample size”-volume)

YBL

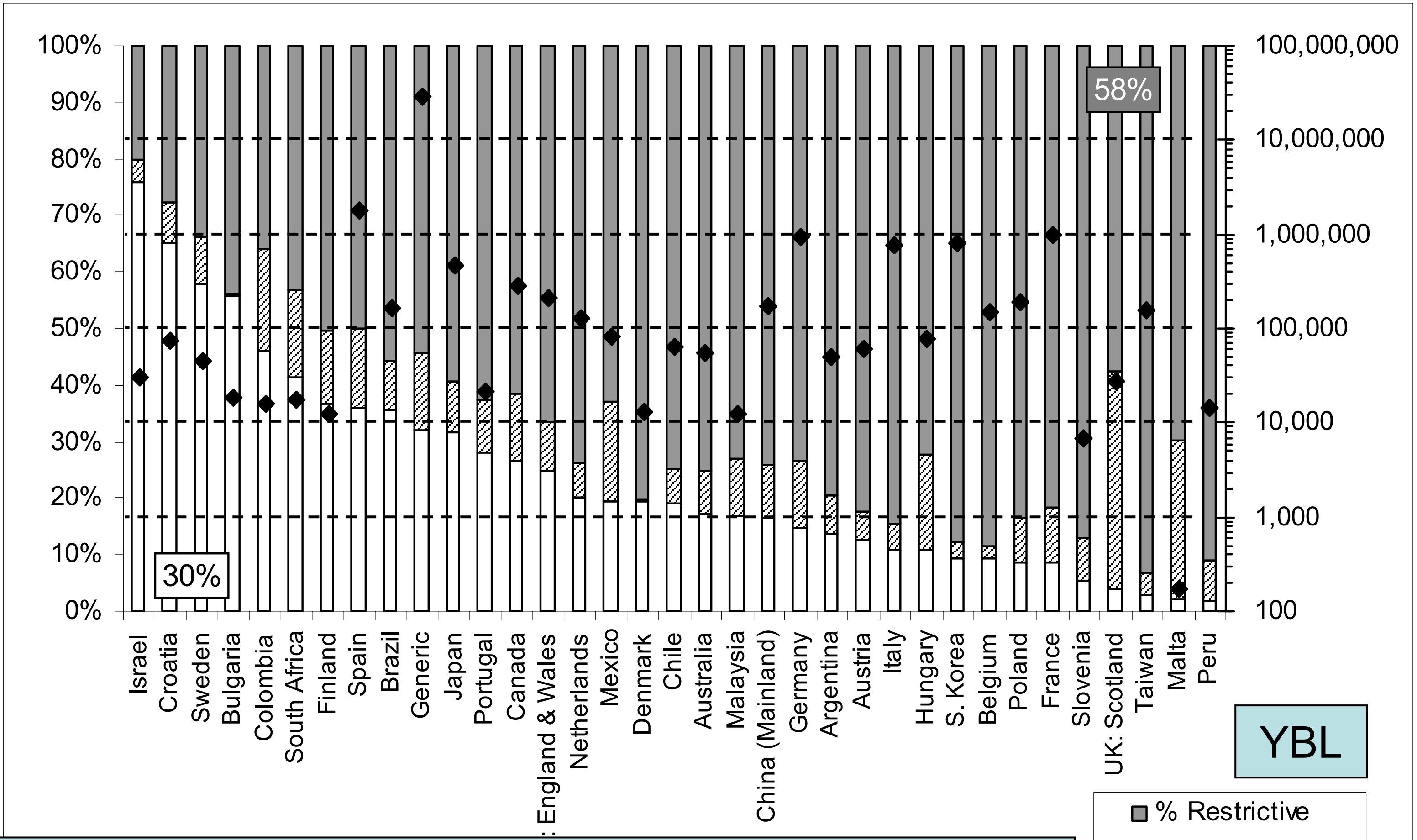
Data presented herein was collected in early 2007. It is based on (imprecise) search engine estimates and is therefore only indicative of the real quantities whose size we are attempting to assess.

Liberal vs. restrictive licensing

- In order to simplify the picture, we can group the 6 licenses into 3 categories: liberal (BY & BY-SA), moderate (BY-ND & BY-NC), and restrictive (BY-NC-SA & BY-NC-ND)
- Then we can sort all jurisdictions according to their relative use of liberal licenses
- Yahoo and Google numbers are not so highly correlated for the license mix *per jurisdiction* as they are for license volume
(in other words, they “agree” more on the number of licensed items *per jurisdiction* than on the license mix *per jurisdiction*)
- However, since our analysis suggests that Yahoo data is more complete, we will use YBL here to compare jurisdictions

Data presented herein was collected in early 2007. It is based on (imprecise) search engine estimates and is therefore only indicative of the real quantities whose size we are attempting to assess.

License mix per jurisdiction (sorted)



- Clear preference for restrictive
- Significant variation, but consistent dislike for moderate licenses
- Jurisdictions with >100k items use >50% restrictive licensing

School of Information Systems

Data presented herein was collected in early 2007. It is based on (imprecise) search engine estimates and is therefore only indicative of the real quantities whose size we are attempting to assess.

SMU SINGAPORE MANAGEMENT UNIVERSITY

iSummit 2007

Freedom ratings to capture "mood"

Proposed license ratings

License	BY	BY-SA	BY-ND	BY-NC	BY-NC-SA	BY-NC-ND
Creative Freedom	6	4	2	5	3	1
Commercial Freedom	6	5	4	3	2	1
Total (Mixed)	12	9	6	8	5	2

Methodology

- Each license is given a freedom rating
- Each jurisdiction is given a rating based on the relative popularity of each license in this jurisdiction
- Optional adjustment for jurisdiction relative volume, to account for the jurisdiction's total contribution to the CC content pool

Data presented herein was collected in early 2007. It is based on (imprecise) search engine estimates and is therefore only indicative of the real quantities whose size we are attempting to assess.

Uses of ratings

- The willingness of the entire CC author population to license their content under more liberal or more restrictive terms can be summarized in just one number, e.g., according to YBL: **6.21** (out of 12)

Freedom rating	Commercial	Creative	Mixed
Generic - YBL	3.38	3.06	6.44
Generic - GBL	3.18	2.89	6.07
All - YBL	3.27	2.94	**6.21**
All - GBL	3.19	2.89	6.08

- Is 6.21 good or bad? Neither, at best what it shows is that the combined effect of the two CC licensing poles (the liberal and the conservative pole) is a rather balanced CC movement, sitting halfway between “all rights reserved” (copyright law) and “no rights reserved” (public domain)
- Interesting is the fact that the commercial freedom values are higher than the creative values. This is because of the popularity of the SA and ND attributes which have a more negative impact on creative freedom than on commercial freedom (according to our definitions)

Data presented herein was collected in early 2007. It is based on (imprecise) search engine estimates and is therefore only indicative of the real quantities whose size we are attempting to assess.

Jurisdiction ratings

- Tables of jurisdiction ratings can be easily constructed for all jurisdictions
- **Jurisdiction ratings should not be hastily interpreted as country ratings!**
  - after all, 80% of the content is under the generic licenses, and this is not only US-based content
  - but ratings are useful as the only global indicator we can automatically construct to assess the willingness of authors in a jurisdiction to license their content under more liberal or more restrictive terms
- Tracking these ratings along with volume data per jurisdiction will allow for some form of measurement of the adoption of the ported licenses in the future

Data presented herein was collected in early 2007. It is based on (imprecise) search engine estimates and is therefore only indicative of the real quantities whose size we are attempting to assess.

Creative freedom ratings (max=6)

YBL

Position	Creative	Rating
1	Sweden	4.2
2	Bulgaria	4.1
3	South Africa	3.8
4	Finland	3.7
5	Spain	3.6
6	Israel	3.6
7	Generic	3.4
8	Brazil	3.4
9	Colombia	3.4
10	Japan	3.3
11	Canada	3.3
12	UK: Scotland	3.3
13	Croatia	3.3
14	Portugal	3.1
15	Poland	3.1
16	UK: England & Wales	3.0
17	Argentina	3.0
18	Chile	2.9
19	Mexico	2.9
20	Netherlands	2.9
21	Germany	2.9
22	Hungary	2.9
23	Australia	2.8
24	China (Mainland)	2.8
25	Austria	2.8
26	Malaysia	2.7
27	Peru	2.6
28	Belgium	2.4
29	France	2.3
30	Italy	2.2
31	Denmark	2.1
32	Slovenia	2.1
33	S. Korea	1.9
34	Taiwan	1.9
35	Malta	1.6

School of Information Systems

Data presented herein was collected in early 2007. It is based on (imprecise) search engine estimates and is therefore only indicative of the real quantities whose size we are attempting to assess.

iSummit 2007

SMU SINGAPORE MANAGEMENT UNIVERSITY

Commercial freedom (max=6)

YBL

Position	Commercial	Rating
1	Israel	4.3
2	Sweden	4.1
3	Croatia	3.9
4	Bulgaria	3.9
5	Colombia	3.7
6	South Africa	3.4
7	Finland	3.3
8	Spain	3.2
9	Brazil	3.1
10	Generic	3.1
11	Japan	3.0
12	Canada	2.9
13	Portugal	2.8
14	UK: England & Wales	2.8
15	Mexico	2.5
16	Netherlands	2.5
17	Chile	2.5
18	Argentina	2.4
19	Australia	2.4
20	Germany	2.4
21	Poland	2.4
22	Malaysia	2.3
23	China (Mainland)	2.3
24	Hungary	2.3
25	UK: Scotland	2.3
26	Austria	2.2
27	Denmark	2.1
28	Malta	2.0
29	Belgium	1.9
30	France	1.9
31	Peru	1.9
32	Italy	1.9
33	S. Korea	1.7
34	Slovenia	1.7
35	Taiwan	1.5

Data presented herein was collected in early 2007. It is based on (imprecise) search engine estimates and is therefore only indicative of the real quantities whose size we are attempting to assess.

Mixed index (max=12)

YBL

Position	Mixed	Rating
1	Sweden	8.4
2	Bulgaria	8.0
3	Israel	7.9
4	South Africa	7.3
5	Croatia	7.2
6	Colombia	7.1
7	Finland	7.1
8	Spain	6.8
9	Brazil	6.5
10	Generic	6.4
11	Japan	6.4
12	Canada	6.2
13	Portugal	5.9
14	UK: England & Wales	5.8
15	UK: Scotland	5.6
16	Mexico	5.5
17	Argentina	5.5
18	Netherlands	5.4
19	Poland	5.4
20	Chile	5.4
21	Germany	5.3
22	Australia	5.2
23	Hungary	5.2
24	China (Mainland)	5.1
25	Malaysia	5.1
26	Austria	5.0
27	Peru	4.5
28	Belgium	4.3
29	France	4.2
30	Denmark	4.2
31	Italy	4.1
32	Slovenia	3.8
33	S. Korea	3.7
34	Malta	3.6
35	Taiwan	3.4

Data presented herein was collected in early 2007. It is based on (imprecise) search engine estimates and is therefore only indicative of the real quantities whose size we are attempting to assess.

Volume-adjusted mixed index

YBL

Position	Mixed	Rating
1	Sweden	8.2
2	Spain	8.2
3	Bulgaria	7.8
4	Israel	7.7
5	South Africa	7.1
6	Croatia	7.1
7	Colombia	6.9
8	Finland	6.9
9	Japan	6.5
10	Generic	6.4
11	Brazil	6.4
12	Canada	6.3
13	UK: England & Wales	5.8
14	Portugal	5.8
15	Germany	5.8
16	UK: Scotland	5.4
17	Poland	5.4
18	Mexico	5.4

Position	Mixed	Rating
19	Netherlands	5.4
20	Argentina	5.3
21	Chile	5.3
22	Australia	5.1
23	China (Mainland)	5.1
24	Hungary	5.1
25	Malaysia	4.9
26	Austria	4.9
27	France	4.6
28	Peru	4.4
29	Italy	4.3
30	Belgium	4.3
31	Denmark	4.1
32	S. Korea	3.9
33	Slovenia	3.6
34	Malta	3.5
35	Taiwan	3.4

Data presented herein was collected in early 2007. It is based on (imprecise) search engine estimates and is therefore only indicative of the real quantities whose size we are attempting to assess.

Looking for relationships…

The differences in the license mix between jurisdictions appear to be unrelated to common economic productivity, political freedom, telecommunications or other national indicators (tested for software piracy level, GDP p.c., unemployment, internet subscribers, broadband penetration, and political, economic and press freedom ratings).

Likely the online communities CC users are active in are the most important determinant of the way they license their content.

But we do observe that…

1. Google and Yahoo jurisdiction data are positively correlated, with volume data per jurisdiction being more strongly correlated than license mix
2. CC has been propelled forward mostly by developed countries with economic, political and press freedom
3. If we examine the top countries in terms of GDP p.c. then only for those countries CC adoption is positively correlated with piracy rates (further study required)

Data presented herein was collected in early 2007. It is based on (imprecise) search engine estimates and is therefore only indicative of the real quantities whose size we are attempting to assess.

Conclusions on CC

License mix

- Authors prefer the most liberal and most restrictive licenses, moderate licenses neglected
- Restrictive licenses significantly more popular than liberal licenses (even if CC users presumably choose CC because they find Copyright Law too restrictive)
- License choice may also depend on the medium type, the community and even the type of content within a medium (ongoing work on these issues)
- Jurisdiction-specific licenses exhibit significant variation from the usage mix of the Generic license

Volume

- The total CC content pool is at least 40-60 million items
- An anti-copyright/pro-piracy attitude may be a strong contributing factor for the growth of CC in some developed economies

Overall

- Belonging to a network/community is probably much more important than belonging to a jurisdiction/country

Data presented herein was collected in early 2007. It is based on (imprecise) search engine estimates and is therefore only indicative of the real quantities whose size we are attempting to assess.

Observations on measuring CC

- Even if we could arrive at some conclusions, the data exhibits significant variations depending on the day of measurement and/or the choice of method
- Search engine results are relatively unreliable for measurement purposes…

  …however by combining several bad measurements we may get a good result!
- Better metadata and proper implementation of CC licensing and search capabilities by search engines and key online communities will be essential for tracking the progress and use of CC

Data presented herein was collected in early 2007. It is based on (imprecise) search engine estimates and is therefore only indicative of the real quantities whose size we are attempting to assess.

If you wish to know more about the study:

giorgos@smu.edu.sg

Thanks to Ankit Guglani, Giri Kumar Tayi, Warren Chik, Anil Samtani, Mike Linksvayer and Lawrence Lessig who helped with producing and/or disseminating this report

Also many thanks to the great folks at the iCommons Summit for their feedback and support

This work is supported by SMU Research Grant 06-C220-SMU-007

Data presented herein was collected in early 2007. It is based on (imprecise) search engine estimates and is therefore only indicative of the real quantities whose size we are attempting to assess.