New Methodologies for Quantifying Licence-Based Commons on the Web

Ben Bildstein
Cyberspace Law and Policy Centre
University of New South Wales
ben.bildstein@student.unsw.edu.au

**First Interdisciplinary Research Workshop on Free Culture,
at iSummit '08, 30 July – 1 August 2008**

Abstract

Current practice in quantifying online commons lacks rigorous methodology. One example of this is in Creative Commons' own data about the growth and state of Creative Commons licensing. Closed, proprietary search engine web services are used to gather approximations to counts of web pages that link to licence URLs. While this data is a useful starting point to get some feel for the current state of Creative Commons-licensed works online, the methodology makes many implicit assumptions. These include: that every licensed work links to a Creative Commons licence; that non-licensed works do not link to the licences; and that proprietary search engines are capable of providing reliable data on links to arbitrary URLs.

This methodology fails to capture some licensed works - works that can have valid plain English licence statements, proper embedded RDF metadata and the appropriate Creative Commons licence mark, but simply fail to link to the licence URL. Nor does this initial methodology generalise well to other categories of documents in the broader commons, such as free software licences or un-annotated public domain works, where the mechanism that creates the public rights is not a link to a URL.

Analysis of the commons as a body of reusable documents, or analysis of the success of the commons movement, requires reasonable data: data about which licences are being used, which ones are most popular in the current environment, and how different media (including image, text, sound, software and others) compare in the make-up of the commons.

This paper proposes using raw web crawler data to do analysis with a reliable methodology. Preliminary experiments and analysis are performed with the purpose of contrast with existing quantification methodologies. Methodological issues about online commons quantification are raised and discussed, including the fundamental methodological question of what constitutes a single creative work on the web: while current practice counts individual web pages (that link to licences), this metric can not easily be applied to media such as motion picture, software, sound and images. Without such a discussion, any data would have only indicative value.

The paper concludes with a discussion of the many areas of potential future work in the quantification of online commons - from deep web and OAI-PMH-compliant databases, to embedded RDF metadata and compressed files, to copies of the full text of licenses as part of the licensed work.

Acknowledgements

The author would like to thank the National Library of Australia for their significant contributions of time, access to data, and computing resources. In particular, thanks to Alex Osborne, who has been particularly helpful in explaining how the network, data access, data formats, and even how some basic Linux concepts work.

1 Introduction

While the software commons movement is now mature, the commons of general web pages is still quite young. The digital commons, as a movement, and fuelled by a lot of enthusiasm for sharing, clearly provides a desire to quantify this emerging commons of creative works. Unfortunately, in the present environment, the tools for quantifying the commons are few, and not very advanced. The major two methodologies for gathering quantification data are as follows.

The first is to collect lists of known repositories of commons documents, and their sizes. Examples of such repositories include licensed wikis (e.g. WikiTravel), web sites of digitised public domain literary works (e.g. Project Gutenburg), or Creative Commons portals to the licensed parts of otherwise unlicensed repositories (e.g. Flickr). The advantage of such an approach is that the figures for individual repositories will be *accurate*, because the repositories have good information about their contents. The disadvantage is that it requires *a priori* knowledge of the existence of the repositories. Nevertheless, this approach is appropriate for providing a known *minimum* to the number of web documents in the commons.

The second current metric for quantification is *backlinks* to licences. That is, measuring the number of web pages that link to appropriate licences. What makes this metric appealing is that it is very easy to collect such data, using proprietary search engines (such as Google or Yahoo Search). For any given licence URL, it is a simple, single query to find approximately how many web pages in the search engine's cache link to the licence. The advantage of this approach is clearly the ease with which data can be gathered. However, the disadvantages are significant, and will be covered in Section 2.

This paper addresses the issues involved with quantifying licence-based commons on the web. It doesn't consider non-licence-based commons, such as the narrow-sense public domain (i.e. works not covered by copyright), because by their very nature there is not necessarily any way to know that such works are in the commons (i.e. there is no obligation to assert the public domain status of a document when publishing it on the web). The focus here is also on HTML web pages, rather than other web-published documents. This means that images, sound, software, etc., are not the main focus.

Also, this is a preliminary, rather than a comprehensive study. In the experiments presented here, not all possible methods have been employed to identify all classes of commons. The focus is on comparison of various methods, with deeper analysis to follow.

Section 2 provides more background to the current practice in quantification of online commons, primarily on the failings of search-engine-based backlinks.

Section 3 proposes alternative quantification methodologies and metrics, based on a raw cache of web documents such as the those used by proprietary search engines to create their indexes and answer queries such as licence backlink queries. These proposed methods are necessarily more resource intensive, and less simple to implement, than the default methodology.

Section 4 describes a cache of web documents held by the National Library of Australia, which is appropriate for use for testing the methods presented in Section 3.

Section 5 details preliminary experiments run on this data of the National Library of Australia.

In Section 6, the results of these experiments are presented; in Section 7, conclusions are drawn from these results, and in Section 8 there is a discussion of the possibilities for future work in this area.

2 Quantification of Online Commons

The discussion and methods of research into quantification of online commons should generally be independent of the precise definition of 'commons'. In fact, such a precise definition is particularly difficult to achieve. For the purposes of this paper, I will adopt an intentionally over-broad definition of commons, made up of cut-down points from the Open Knowledge Definition v1.0:[1]

1. Access: The work shall be available as a whole and at no more than a reasonable reproduction cost, preferably downloading via the Internet without charge.[2]
2. Redistribution: The license shall not restrict any party from selling or giving away the work either on its own or as part of a package made from works from many different sources. The license shall not require a royalty or other fee for such sale or distribution.
9. Distribution of License: The rights attached to the work must apply to all to whom the work is redistributed without the need for execution of an additional license by those parties.

This working definition of commons includes the public domain, free[3] and open source[4] software, Creative Commons licensed works,[5] and many others. But it will also include licences that some would consider non-free or non-open, such as the NonCommercial Creative Commons licences.

As described in Section 1, the two main sources of quantitative data about the commons in current practice are: known repositories (or content directories), and search engine backlinks. Some of the online commons will be in each, some in both (search engine indexed web pages that are also listed in commons content repositories), and some in neither (unknown repositories of commons content that are not search engine indexable).

Content Directories

Content directories, or repositories, are arguably the most reliable for identification and quantification, because the licence-based metadata is recorded on a per-document basis. That it, it is a human and not a computer that is *identifying* the documents as belonging to the commons. But, as described in Section 1, this method can not provide an accurate size to (any part of) the commons, because there is no way to systematically find all appropriate repositories.

Search Engine Backlinks

Proprietary search engines are opaque by nature; details of the underlying indexes of web content used for performing searches are not publicly available, as the technology and quality of the indexes are a competitive edge for the search engine company. This means that the raw numbers provided by search engine backlink quantification are difficult to interpret.

One problem is that the size and coverage of the index is unknown. For example, does the search engine limit its indexing to a maximum number of pages for each domain? Or is it limited to a certain depth, such that for a page to be indexed it must be reachable in a certain number of 'clicks' from the top level page?

---

1 *The Open Knowledge Definition*, The Open Knowledge Foundation, <http://opendefinition.org/1.0/>
2 The original includes: “The work must also be available in a convenient and modifiable form,” which is a reference to the 'open standards' issue.
3 *The Free Software Definition*, The Free Software Foundation, <http://www.fsf.org/licensing/essays/free-sw.html>
4 *The Open Source Definition*, Open Source Initiative, <http://www.opensource.org/docs/osd>
5 *Creative Commons Licenses*, Creative Commons, <http://creativecommons.org/licenses/>

Dynamically generated content

Another problem is that there may be pages in the search engine's index that do not actually have content. For example, web server error pages such as 'document not found' pages may be included in the index. This might happen due to broken links in the web, where a document's location changes, but there is still a link to the old location, the search engine finds the old link, requests the invalid page, and receives an error message that it is unable to distinguish from otherwise useful content.

Another example of contentless documents indexable by search engines is dynamically generated web pages. For example, a web site might have an online calendar of upcoming events, where a user can navigate to a specific URL for any given date, past, present, or future. The web server can have a database of events, and dynamically produce a 'what's on' page for any date, filling it out with any events from the database. In this case, though there may be only one or two events in the database, almost limitless pages 'exist' (in the sense that they are accessible and indexable, if not actually created and stored in advance), with most having the same content, that being a statement that there are no events entered in the calendar for the given day.

Another type of dynamically generated web pages that don't contain unique content is site-specific search results. Here, a web server might have a database of records of some kind, and a search interface for finding records. When a search is performed, a web page of search results is generated, and this page can have its own URL, and hence can be indexed by search engines. Thus, any number of queries can be turned into entries in the search engine's index, yet the content does not map to the actual entries in the web site's database, but rather to arbitrary and overlapping lists of entries.

These types of dynamically generated web pages arguably do not contribute anything to the online commons, in terms of documents that people would find useful to reuse. Additionally, the number of such dynamically generated pages could in fact be quite large; in each example, there is essentially no limit to the number of pages that could be generated.

Other problems with backlinks

One of the problems with using backlinks for quantification goes beyond the opacity of search engines' indexes, and is fundamental to the concept of a backlink: a link to a licence does not necessarily mean that the linking web page is being offered under that licence. Online discussion of the licences leads to web pages linking to them, and these two types of backlinks, first where the licence is *used*, and second where the licence is *mentioned*, are indistinguishable when simply counting backlinks.

On the other hand, a web page may be offered under a licence yet not link to the relevant URL for that licence. In fact, arguably the most meaningful statement that a web page is available under a licence is a written statement. In the Creative Commons system, the statement would be something similar to the following: “This work is licensed under a Creative Commons Attribution 2.5 Australia License.”[6] Such a statement might not include a link to the relevant[7] Creative Commons licence URL, yet such web pages may still be appropriate for inclusion in the online commons.

There are also other licensing *mechanisms* (that is, ways of stating that a document is available under a licence) that backlink quantification will miss. These include the *rel-licence microformat*,[8] *RDF+XML / RDFa*, and *HTML meta tags*. These mechanisms will be considered further in Section 3.

---

6 From <http://creativecommons.org/license/>

7 In this case, that would be <http://creativecommons.org/licenses/by/2.5/au/>

8 *rel-license – Microformats*, Microformats Wiki, <http://microformats.org/wiki/rel-license>

Backlinks for licence comparison

The opacity of proprietary search engines means that the precise numbers reported by them are of unknown accuracy.[9] However, it may still be reasonable to consider the search engine to be unbiased in many respects. If a search engine is used to give an approximation of backlink numbers for two licence URLs, while the absolute numbers reported might not represent the number of actual licensed web pages that use them, assuming reasonable precision, if not accuracy, the relative size of the backlink counts, with respect to each other, may be meaningful [3]. This technique can be useful measuring the comparative popularity of various licence attributes, such as Creative Commons' Attribution, No-Derivatives, NonCommercial and ShareAlike.

Additionally, this technique may be useful in monitoring relative changes in the size of the commons over time, under the assumption that the precision of the search engine's index is not variable over time. That is, under the assumption that the ratio of backlink numbers reported by search engines to the actual size of the online commons (or the parts specific to particular licence) is constant over time, it would be possible to report accurately on the growth of the online commons (or specific parts of the online commons) over time.

9 As discussed in [4], they might tell us accurately the size of the index, but not the size of the underlying web space.

3 Alternative methodologies for quantifying online commons

As described in the previous section, the fundamental problems of the search engine backlink quantification methodology are:

1. Opacity: the criteria for inclusion in the index are unknown, so data may be biased in unknown ways.
2. Backlinks: a backlink is not a licence offer, so web pages will be picked up that should not be counted.
3. Dynamically generated content: search engine backlink numbers will count dynamically generated pages that may not be appropriate for inclusion in the online commons, in terms of reusability.

This section addresses these issues and proposes alternative methodologies.

What is a unit of the commons?

At this point it is necessary to discuss a fundamental issue in the quantification of commons, whether online or offline. The search engine backlink methodology has become a *de facto* standard. There are a few possible reasons for this. First, it is easy to implement. Second, there are no obvious alternatives (with the possible exception of repositories, as described in Section 1). Third, the results support the goals of those performing the quantification. That is, the numbers are very large, and as such support the argument that the movement towards a global commons of free culture has significant momentum and should not be ignored.

The reason the numbers are large is because they come from counting URLs (web addresses). The URL has been chosen as the implied unit of measurement of online commons. This is a dubious decision (even if not taken consciously), for a number of reasons. As discussed, dynamically generated content implies that a unique URL does not necessarily map to a unique document on a web server. When not being viewed, in a very real sense the content may not actually exist. Or, multiple URLs may map to the same actual file (on the web server), and so when viewed may be identical.

Second, a web page has an arbitrary amount of content in it. No comparison can be made with other media in terms of size. For example, how much does a web page contribute to the commons compared to a published book? The fact is that we have a reasonable idea of the size of a book. If we say 10,000 published books, this is a meaningful amount, because there is realistically a minimum and maximum length that a book might be. Similarly, a number of images is meaningful, because they are all approximately the same size when viewed (large pictures can be scaled down and vice-versa). Music can reasonably be measured in seconds or minutes of audio.

This paper does not aim to make a statement about the *value* of works that are potentially in the commons, but it is very important to be clear on exactly what is being measured. If we measure URLs, the data will be biased towards pages that are accessible via multiple URLs. If we measure web pages (if that is indeed possible), the data will still be biased towards documents that are cut up in to many pieces (compare, for example, a book that is available online as a single long web page with a similar book that has each page of the book available online as its own web page). Yet if we choose to measure Internet domains, as will be discussed later, results are biased towards domains with less content.

Crawler based approach

The obvious way around the problem of search engine opacity is to crawl the web to create a distinct and custom web cache similar to those held by search engines, or to otherwise obtain access to such a web cache. Then questions can be answered directly and in depth from the web data, rather than in the summary form provided by search engines.

An additional, and significant benefit is that other licensing mechanisms can be used, rather than simply backlinks. For example, embedded licensing metadata can be identified.

One other advantage of this approach is that it allows the licensed web pages to be identified. While search engines will give approximate numbers of licensed pages, and it is possible to view some of these pages (i.e. the search results), there is usually a limit to the number of search results that can be viewed. For example, using the Google search engine, only the first 1,000 results are viewable. On the other hand, with a custom web crawler approach, all web pages identified could be enumerated (for example to create a database of online commons).

A first alternative to counting URLs

One conceptual alternative to counting URLs is count *people choosing to apply licences to their content*. In terms of analysing the *movement* towards an online commons, this could yield valuable insight into the participants of the movement. On the other hand, in terms of knowing how much content is available to reuse it may have limited application.

To illustrate this idea of *distinct uses* of the licences, consider a blogger who chooses to licence their blog under an open content licence. The licence statement, logo and hyperlink can be added to the template for the blog such that it shows up on every page, with the intention that all pages, both pre-existing and yet-to-be-written are available under the licence. In this case, we have many web pages or URLs that count as part of the online commons, yet only one blog, or only one *distinct use* of the licence. It is both interesting and relevant to be able to analyse the relationship between licensed web pages and distinct uses.

Unfortunately, there is no clearly best way to decide, by looking at the licensed web pages, what is and is not a distinct use. In fact, the best way would be to interview web page owners on a page by page basis, though this is clearly infeasible.

What is presented here are a few heuristics, as a starting point for broader research into online commons quantification.

Grouping URLs into logical web sites

Although there are many ways we could choose to count licensed documents, one way that provides an interesting contrast to counting URLs is to count web sites.[10] Again, there is no simple (automatable) way to group web pages into web sites, but this paper presents a few heuristics.

Distinct domain names

The simplest grouping is to group web pages that have the same domain name in their URL. For example, the following two URLs, which share the same domain name, would be grouped together: *http://en.wikipedia.org/wiki/Go_(board_game)* and *http://en.wikipedia.org/wiki/Edsger_W._Dijkstra*. Yet *http://nl.wikipedia.org/wiki/Edsger_Dijkstra* (the Dijkstra page from the Dutch Wikipedia) would not be grouped with the former two. So this

10 “A location connected to the Internet that maintains one or more pages on the World Wide Web”, *The Australian Concise Oxford Dictionary* p1625, Oxford University Press, 2004

grouping would assign to each language's Wikipedia it's own group. Already, we can see that the arbitrary decision (by those that administer Wikipedia) to give each language its own subdomain, has had an effect on this heuristic's groupings.

Connected components

The next possibility is that if one page links to another page, they could be considered part of the same group. This would almost certainly mean that a wholly licensed web site could be identified as such. The drawback is that a link between two licensed web sites would in fact draw them together into the same group. As such, the effectiveness of this heuristic is based upon the assumption that there are few links between licensed sites, or at the very least, that the licensed part of the web is small. Although this may seem a significant drawback, as the results section below will show, this effect is rather smaller than might be expected.

At this point it is worth taking a moment to consider the technical implementations of what may constitute a 'link' between two pages. Of course the obvious answer is a hyperlink: either page has a direct hyperlink to the other. Other options include (ranging broadly from weaker connection to strong connection):

- There is a path using at most N hyperlinks, from one page to another, where N is some number;
- There is such a path in both directions;
- There is a path (using multiple hyperlinks) from one page to the other;
- There is such a path in both directions;
- Both pages have direct links to each other.

The ideas of grouping based on linking and grouping based on domain name can be combined to address the issue of domains that host multiple web sites. The advantage of this is that it mitigates the effect of *percolation* [10], which is that when the average node of a network such as the World Wide Web gets sufficiently well connected to other nodes, percolation occurs and nodes of arbitrary distance apart will be connected. In this area, this corresponds to the *licensed web* becoming a large enough portion of the World Wide Web that there a considerable chance that there is a path, made up only of links between licensed pages, between any two (otherwise unrelated) licensed pages.

4 The Australian web domain harvests

In each of 2005, 2006 and 2007, the National Library of Australia commissioned a crawl of Australian web documents for archival purposes, and it is likely to do further crawls in the future. These crawls were each outsourced to the Internet Archive, who ran the crawls and delivered the relevant hardware, populated with the results of the crawls of the Australian web. An explanation and preliminary quantitative analysis of the data is presented in [7].

Web pages were included or not based on the associated domain name and IP address. If the domain name ends in ".au", then it was considered Australian. If not, then the IP address was considered for mapping to a computer located in Australia. A geolocation database, MaxMind, was used for determining the associated country for an IP address.[11] Additionally, the crawls have included web pages that are directly linked from Australian web pages. This means that there are some web pages included that are not Australian web pages.

The scope of the 2005 crawl was to crawl as much as possible within four weeks. The 2006 and 2007 crawls were set to crawl at least 500,000,000 distinct URLs [7].

11 Personal communication with Paul Koerbin and Alex Osborne, see also <http://www.maxmind.com/app/country>

5 Experimental design - Quantification of licensed web pages in the Australian web domain harvests

To investigate the possibilities of the alternatives to backlink quantification discussed in Section 3, a small experiment was run on a subset of the National Library of Australia's 2006 and 2007 data. These years were chosen because they shared the same collection policy, and so a meaningful comparison between the two years' data should be possible.

To limit the amount of analysis required, only pages in the *.edu.au* second level domain were analysed. As such, the results of this analysis will not represent the online commons more generally. In the crawl data, there were 30,888 hosts returning pages from *.edu.au* in 2006, and 31,398 in 2007. 33,369,094 URLs were crawled from *.edu.au* in 2006, and 39,609,285 in 2007. The proportion of the 2006 and 2007 crawls that were in the *.edu.au* second level domain was just under 10% in both cases, second only to *.com.au*. Over the whole years' crawls, 86-87% of requests were retrieved OK (response code 200), with the remaining percentage fairly evenly divided between temporary redirects (code 302) and not found (404).[12]

The analysis of the *.edu.au* second level domain was run during June and July 2008. The major phases of the analysis were: identification of potential commons, grouping, and quantification summary.

Identification of potential commons

Three methods of identifying potential commons (referred to as licensing mechanisms) were implemented. First, links to Creative Commons licences (i.e. backlinks). Ideally, a large database of known commons-based licences would be used, but Creative Commons was used for simplicity's sake.

Second, the *rel-license* microformat,[13] which is simply a statement that the linked-to page holds a licence that the linked-from page is available under. The licence being referred to was not considered, so it is possible that some works may have been identified as potential commons which actually had licences that do not qualify under even the broad definition of commons used here.[14]

Third, Dublin Core "DC.Rights" metadata expressed in HTML meta tags, that referred to Creative Commons licences.[15] Specifically, this looked for meta tags with *name="dc.rights"* allowing upper-case, and a any other non-alphanumeric character instead of the dot (e.g. *DC.Rights* or *dc:rights*). The reason that only Creative Commons licensed works were identified with this mechanism is that any statement of rights is valid in this field – it is not a commons-specific mechanism. For example, it is possible and perfectly reasonable to state using Dublin Core metadata that all rights are reserved (e.g. <*meta name="DC.Rights" content="Copyright Ben Bildstein 2008, All Rights Reserved"*>).

For both backlink and Dublin Core quantification, the following regular expression was used to determine if a URL was a Creative Commons licence: "^http://(www\.|)creativecommons.org/licen[cs]es/." This can be explained in natural language as follows: *the URL starts with "http://", then optionally has the text "www.", then has the text "creativecommons.org/licen", then has a 'c' or an 's', then has the text "es/", then has at least one*

12 [7] at pages 7 and 9
13 See note 8.
14 See Section 3.
15 *Expressing Dublin Core in HTML/XHTML meta and link elements* <http://dublincore.org/documents/dcq-html/>; *Guidelines for implementing Dublin Core in XML* <http://dublincore.org/documents/dc-xml-guidelines/>

more character. This method of deciding whether or not a page is a Creative Commons licence was taken from the source code of CcNutch.[16] [17]

The pages that were tested against the above three criteria were all those in the *.edu.au* second level domain from the 2006 and 2007 crawls that had response codes of 200 OK and MIME types of of text/html or application/xhtml+xml. All Creative Commons licence hyperlinks, rel-license labelled links, and Creative Commons meta tagged Dublin Core rights metadata were recorded.

Grouping

Having identified a collection of documents, the next step was to group them. Of the many possibilities here, the three that were attempted were: grouping pages with the same domain name, grouping pages where one has a link to the other, and grouping pages where one has a link to the other *and* both pages have the same domain name.

The latter two, which will be referred to as grouping into *subwebs*, were performed using a union/find data structure.[18] Terms in bold in the following explanation are defined in [6]. From a graph-theoretic perspective, the whole crawl was considered as a **directed graph** (where URLs form the **vertices** and hyperlinks form the **edges**), and the *licensed web* was the **undirected version** of the **subgraph** that is **induced** by the set of pages identified as possible commons. The two groupings that were performed were then the **connected components** of the licensed web, and the connected components of the licensed web with edges removed where those edges cross a domain name boundary (that is, the pages that the edge connects have different domain names).

Quantification summary

Apart from grouping pages according to each of the three schemes of domain names, subwebs, and domain-isolated subwebs, the results of the commons identification were also summarised into six statistics for each year. These statistics were, for each of the three mechanisms, backlinks, rel-license, and DC.Rights meta tags, how many pages were identified under the mechanism and how many licensing statements of that type were found (i.e. a page may contain more than on link to a Creative Commons licence, more than one rel-license tag, and/or more than one DC.Rights meta tag).

16 CcNutch is the Creative Commons aware module for the open source search engine Nutch. <http://wiki.creativecommons.org/CcNutch>

17 The source code file used was <http://code.creativecommons.org/svnroot/ccnutch/trunk/src/java/org/creativecommons/nutch/CCParseFilter.java>

18 For more on the disjoint set data structure, see <http://en.wikipedia.org/wiki/Disjoint-set_data_structure>

6 Results

The following table summarised the number of licensed pages identified and the mechanisms by which they were identified for both the 2006 and 2007 crawls:

	2006	**2007**
Has <meta> DC.Rights mechanism	282	292
Total <meta> DC.Rights tags	282	292
Has 'rel-license' link mechanism	39,391	60,843
Total rel-license links	53,396	63,892
Has link to CC licence	53,498	143,181
Total links to CC licences	69,215	182,981

The following table shows the total number of groups for each of the three applied grouping options:

	2006	**2007**
Domains	119	139
Disjoint subwebs	445	13,217
Domain-isolated disjoint subwebs	488	13,241

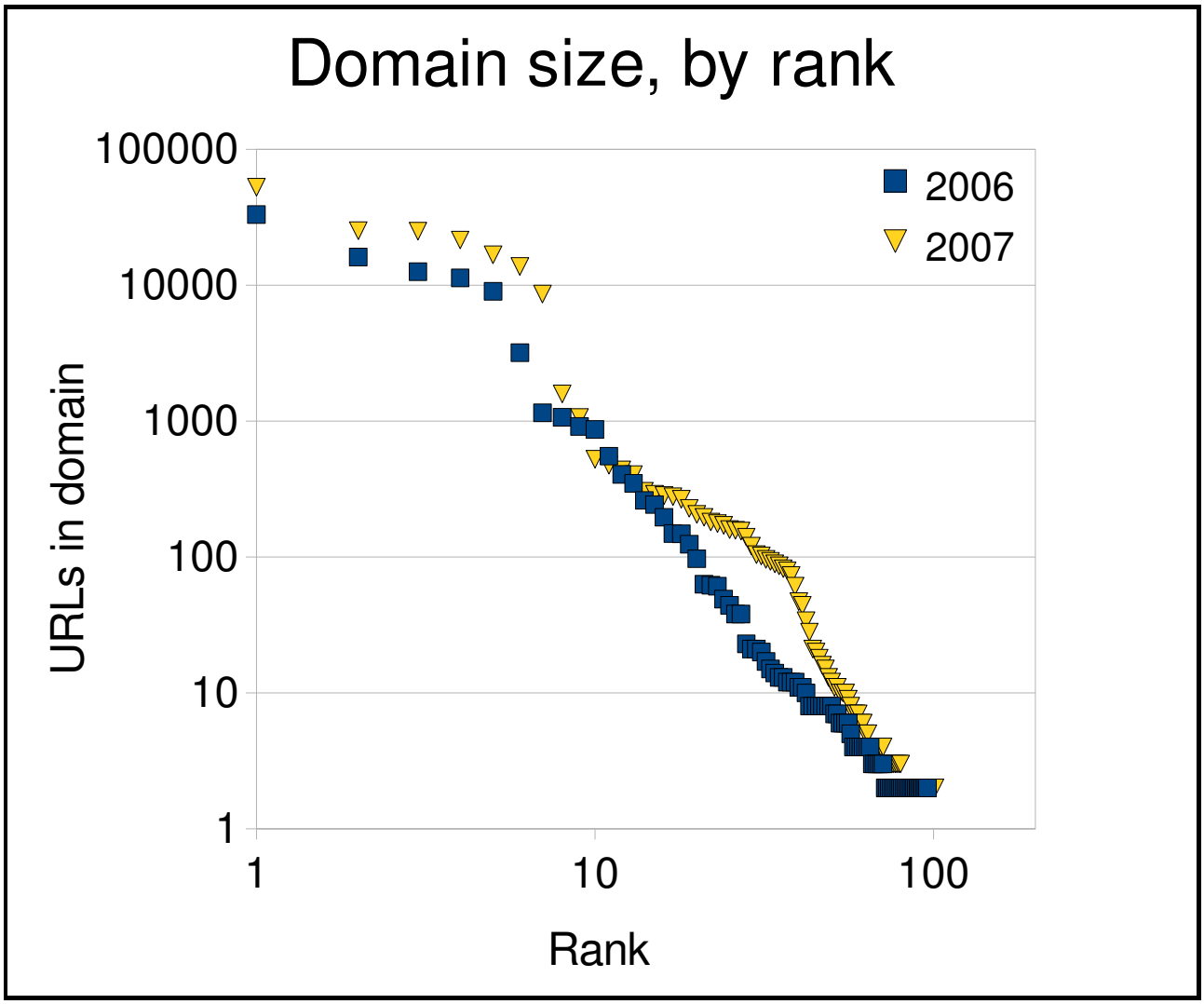
The large number of subwebs for the 2007 data is due to a large number of web pages that did not link to other (cached) pages on the same site. The lack of links between pages on the same site here led to every such page being grouped into its own subweb. There were 12,437 such single-page groups from www92.griffith.edu.au. The largest group of single-page groups from the 2006 data was 52 from www.dssrg.curtin.edu.au/~satherrl/. Clearly, this shows that a lot more work needs to be done before the simple number of groups is a meaningful quantification metric. However, the sizes of the groups can be plotted into a graph, as in the next subsection, and this does indeed turn out to be an interesting visualisation tool.

Graphs

The following five graphs all display the various groups under the various grouping systems. In each case, the groups have been ordered from largest to smallest, and graphed on a logarithmic vs. logarithmic scale. In each of the graphs, only groups with a size of two or more are plotted.

Grouping by domain only

This first graph shows the results of grouping based only on the domain name, and shows the two years of data as separate series.



The first thing to recognise from this graph is the power law relationship between the series. Plotted without the logarithmic axes, the scale of the graph would be entirely dominated by the first few very large points, and the very long tail of very small domains.

Visually, there are three common features between the 2006 and 2007 series. The largest 5-7 domains do not have the same slope as the graph as a whole. But after these first few points, there is a discontinuity: the number of licensed works per domain drops by a factor of almost 1000 in just 1-2 points.

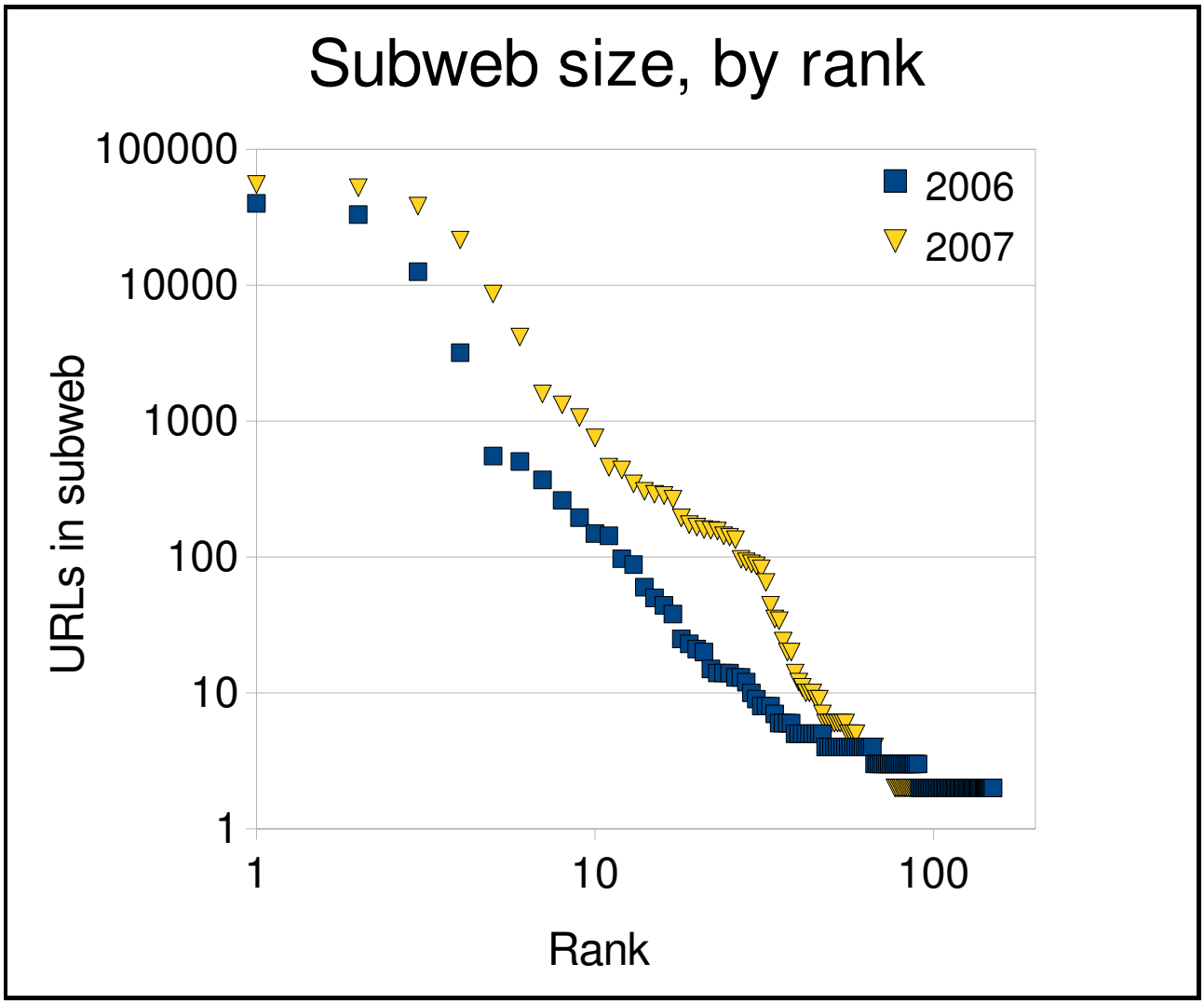
Because this flat area and discontinuity happened around the area of the largest sites, there are two clear possibilities for why this happens. First, they may be an artefact of the crawl parameters: it is a common idea in web crawling to put a limit on the number of pages to be downloaded from a single domain, so it may be that the underlying reality of these sites is that they are actually much larger than represented here.

The second possibility is that the shape of this graph represents the make-up of the commons; that there truly are two distinct parts to this graph. If this is so, the mechanism of exploration for this effect is unknown.

Thirdly, the two series follow different paths around the 25th point. The explanation for this might be that wholly licensed web sites have grown between 2006 and 2007.

Grouping by connected components

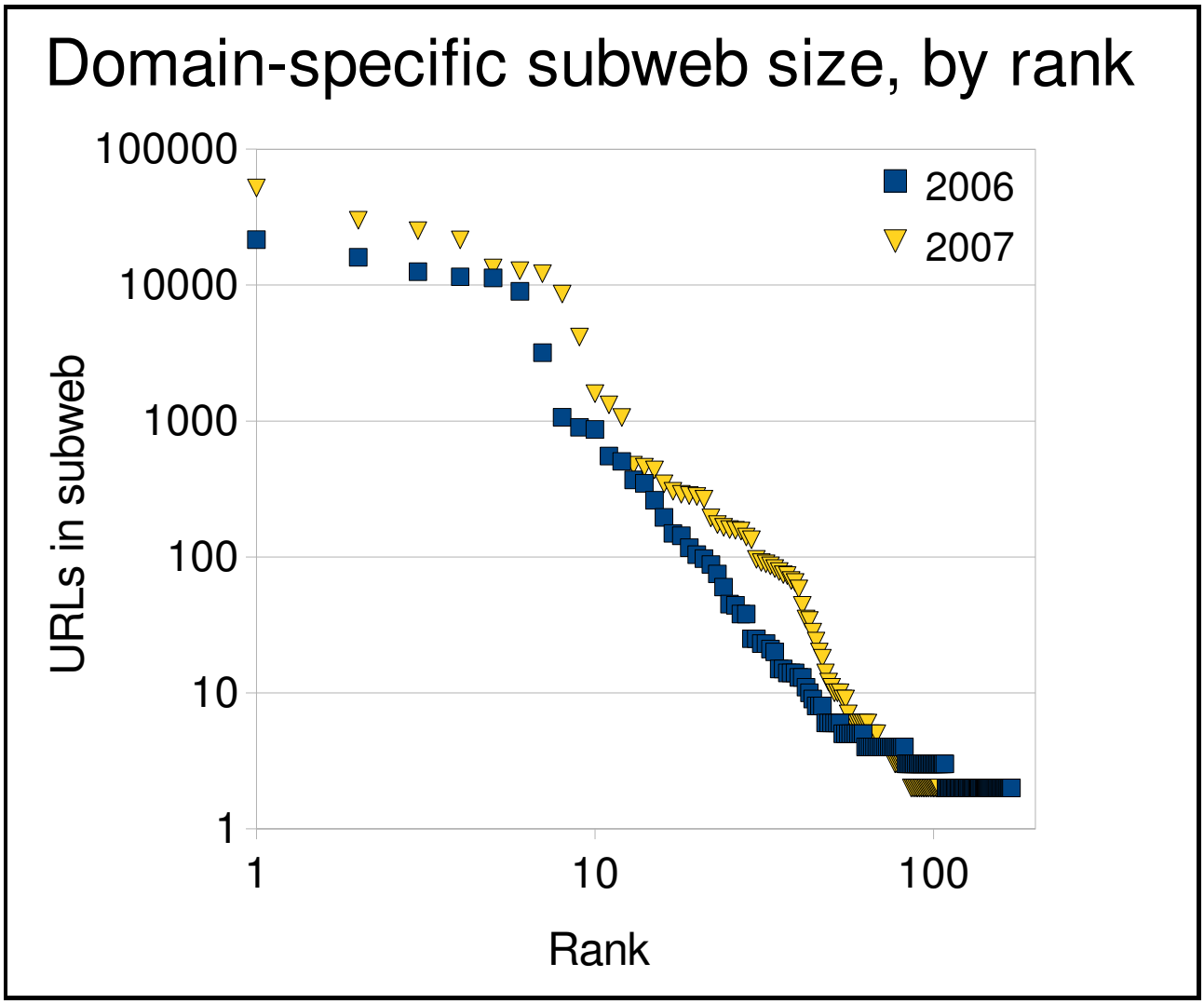
This graph shows essentially the same thing for subweb analysis: pages are in the same group if and only if there is a hyperlink path between them via only licensed pages.



Compared to the previous graph, this one is more continuous (i.e. smoother). There is still a marked change in the nature of the graph the the 5th data point of the 2006 data and the 7th data point of the 2007 data, but the points leading up to this change are more continuous. Again, the 2007 data lies generally higher than the 2006 data, which suggests that over the year, either licensed web sites got bigger, or new, bigger, sites were either licensed or created.

Grouping by domain-isolated connected components

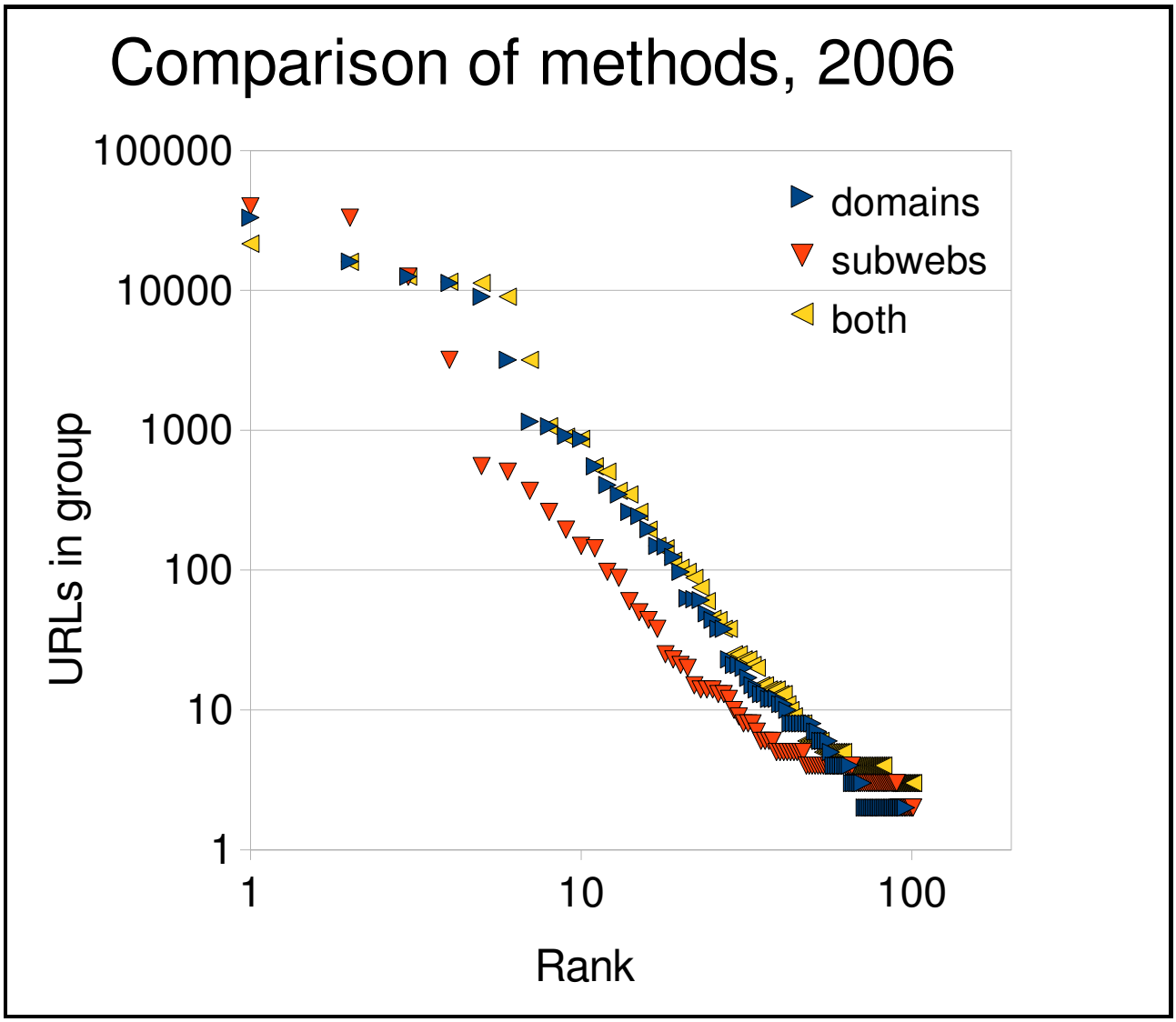
This third graph shows groups where two pages are in the same group only when they have the same domain name, and there is a path between them as in the previous graph.



This graph shows basically the same features as the previous two graphs, but a detailed comparison will be left for the graphs that compare the various grouping methods directly.

Comparison of grouping methods, 2006 data

This graph plots the data from all three grouping methods, for the pages identified as potential commons in the 2006 data. The series are the same as from the previous graphs of 2006 data.



The most interesting observation from this graph is that the 'domains' series and the 'subwebs' series are remarkably similar in shape. In fact, there seems to be more similarity between these two independent grouping methods in this graph than there does between different years of data with the same grouping method (the previous three graphs). And note that *a priori*, there is no fundamental connection between the 'domains' groupings and the 'subwebs' groupings. This similarity suggests that both of these types of groupings are good at isolating the underlying licensed web *sites* (collections of web pages), and secondly that such web sites tend to have only a single domain name.

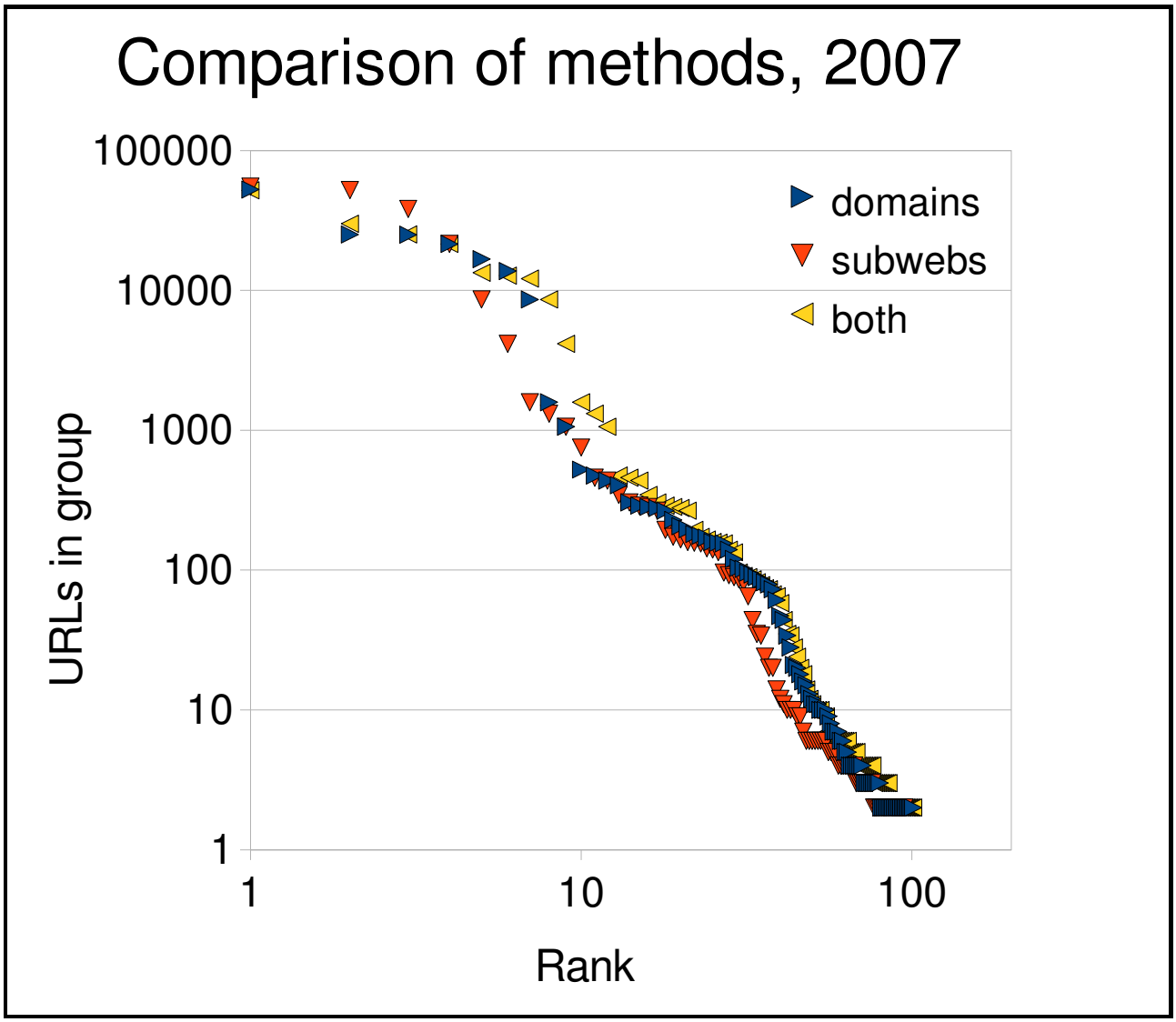
The 'subwebs' series appears to be similar to the others only shifted to the left. This is due to combining groups (data points) that are distinct in the other two series because they exist on multiple domains.

The third datapoint is the domain epinet.anu.edu.au, where all three series align perfectly in both dimensions, with 12,561 potentially commons pages. However, note that other such wholly-domain-contained subwebs can be spread out on the graph: the three points lying directly between the 1,000 and 10,000 lines correspond to a group of 3,182 pages from the domain www.woodvale.wa.edu.au.

The fact that there is very little horizontal shift between the 'domains' and the 'both' series suggests that it is uncommon for a domain to hold multiple subwebs.

Comparison of grouping methods, 2007 data

This graph plots the data from all three grouping methods, for the pages identified as potential commons in the 2007 data.



Again, in this graph we see that the three series have remarkably similar shape, suggesting that the various grouping methods are relatively good at identifying web sites in the licensed web.

The fourth datapoint, where all three series align, is the domain apsa.anu.edu.au, with 21,487 pages.

In terms of overall shape, there are four main features of this graph: a relatively flat part (up to about the 7th point, at about 10,000 URLs per group), followed by a relatively steep or even discontinuous part (to the 10th point, 400 URLs), then a relatively shallow part (to the 30th point, 100 URLs), then a quite steep decline towards the tail. This reason for this overall shape is unknown and requires further investigation.

7 Conclusions

The body of commons on the web is very complex in its nature and make up. Firstly, the concept of quantification is complex. Simply put, what should we be counting? There are clear examples of why it is not appropriate to rely on backlinks as the complete answer: parts of the licensed web that are large in terms of number of URLs, but have very little content.

The grouping explored here also shows that there is no typical licensed web site. Most sites that have licensed content (at least in the *.edu.au* domain) have only a handful of licensed pages, yet the vast majority of licensed URLs are found in a very small minority of sites.

A full understanding of the dimensions of the commons is still likely to take a while. This paper has shown the complexities of the problem, and demonstrated some initial heuristics for, and possibilities for future research into, dealing with these issues.

8 Future work

The research presented here is definitely work in progress. Many aspects of it require more consideration and more validation, and many potential methodological options were not fully explored.

Scale up

The most obvious limitation of this research is its small scale. It could be scaled up in three ways. First, more of the National Library's data could be used: 2005 data as well as the 2006 and 2007 data; or data from the rest of the crawl instead of just *.edu.au*. Even without increasing the analysis time required, this could be done with some suitable random sampling of the data.

The second option is to find a larger data source. The Internet Archive is an obvious candidate, as they are continuously crawling and archiving the web.[19] Another candidate is Wikia Search,[20] which uses a distributed web crawler called Grub to crawl the web continuously.[21] The crawl data from Wikia Search is available online.[22]

A third option is a dedicated crawler. While this would require significant resources, it has the advantage of giving more control over crawl parameters, and more transparency in the data.

Media

This paper has primarily dealt with the commons of web pages. Yet there are also commons of images, sound files, video, software and other media that must be investigated before a full understanding of the commons can be achieved. The idea of graphing items of the commons based on their size will make for an interesting comparison between media. Some options include ranking sound files by length, or software products by lines of code. Much more investigation of such options is needed. Yet such investigation will be valuable. For example, it would be no use to know the number of free or open source software products in existence if most of them are barely started and functionally useless.

Mechanisms

In terms of licensed web pages, more work needs to go in to the identification of licensed content. The examples include RDF+XML, where licensing statements are expressed in XML and then embedded in HTML comments, and automatic recognition of full-text copies of licences, which can be used exactly like the originals (for example, a link to a full-text copy of a licence may indicate licensing).

Beyond crawling

More exploration is needed of the amount of commons content that is not indexable by crawlers. This would include content in databases such as Flickr, deep web databases (where content is not accessible via hyperlinks from the main web, but rather through submission of an HTML form from the main web), or even OAI-PMH compliant databases (The Open Archive Initiative's Protocol for Metadata Handling – a standardised way to gather metadata such as intellectual property rights from

19 *Internet Archive Web Archive*, <http://wa.archive.org/>
20 <http://search.wikia.com/>
21 *Crawl the Web*, Wikia, <http://re.search.wikia.com/about/crawl.html>
22 <http://soap.grub.org/arcs/>

databases on the Internet).

Qualitative studies

The issues of what to count when measuring the commons, and how to discount documents that have no real content, blur the line between quantitative and qualitative research. To find the most meaningful way of quantifying the commons, it may be necessary to do case studies of commons publishing on the web, to try to get a feel either for what is typical, or if there is no typical publishing, what the spectrum of publishing encompasses. For example, the site www.anu.edu.au/people/Roger.Clarke/ contains pages that have been licensed individually. That is, every article on the site that is licensed has been hand-chosen by the author of the site for licensing. This means that in a quantification, every licensed page on the site has real content, unlike other sites which have licences applied to pages that have no real content because the licence is applied to every page within the site as a policy.

9 References

[1] Adida, B., Birbeck, M. (editors), *RDFa Primer*, working draft, 20 June 2008, <http://www.w3.org/TR/2008/WD-xhtml-rdfa-primer-20080620/>

[2] Broder, A., Kumar, R., Maghoul, F., Raghaven, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J., *Graph structure in the web,* Proceedings of the Ninth International World Wide Web Conference, 2000, <http://www9.org/w9cdrom/160/160.html>

[3] Cheliotis, G., *Creative Commons Statistics from the CC-Monitor Project*, 2007, <http://hoikoinoi.files.wordpress.com/2007/07/cc-monitor-findings-isummit.pdf>

[4] Cheliotis, G., Chik, W., Guglani, A., Tayi, G., *Taking Stock of the Creative Commons Experiment: Monitoring the Use of Creative Commons Licenses and Evaluating Its Implications for the Future of Creative Commons and for the Copyright Law*, 2007, 35th Research Conference on Communication, Information and Internet Policy (TPRC)

[5] Cheliotis, G., Guglani, A., Tayi, G., *Measuring the Creative Commons* (extended abstract), 2007, submitted to the Third Symposium on Statistical Challenges in E-commerce Research, <http://www.citi.uconn.edu/scecr07/content/CheliotisGuglaniTayi.pdf>

[6] Cormen, T., Leiserson, C., Rivest, R., *Introduction to Algorithms,* 1990, The MIT Press

[7] Koerbin, P., *The Australian web domain harvests: a preliminary quantitative analysis of the archive data*, 2008, National Library of Australia, <http://pandora.nla.gov.au/documents/auscrawls.pdf>

[8] Kumar, R., Raghaven, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., Upfal, E., *Stochastic models for the web graph*, 2000, 41st Annual Symposium on Foundations of Computer Science

[9] Lessig, L., *Cyberspace's Architectural Constitution*, speech to the Ninth International World Wide Web Conference, 2000, <http://cyber.law.harvard.edu/works/lessig/www9.pdf>

[10] Weisstein, E., "Percolation Theory", from *Wolfram MathWorld*, <http://mathworld.wolfram.com/PercolationTheory.html>

[11] "License Statistics", *Creative Commons Wiki,* <http://wiki.creativecommons.org/License_statistics>