The Unit of the Commons

- A discussion of metrics for online commons quantification
Why quantify?

- Justifying commons theory as a field of research
- Informing commons theory
  - Regulation of public goods
  - The commons in economics
- Informing licensing organisations
- Informing law and policy makers
Media

- Text
- Images
- Sounds
- Video
- Software
Overview of current quantification results

- Creative Commons: CC wiki says '130 million total works'
- FOSS: SourceForge has 'more than 230,000 software projects'
- GFDL: Wikipedia has 2.83 million articles in English
- Photos: 102,868,725 photos
Taken from: 2005 - Dormida (http://www.flickr.com/photos/nfjoral/3444556932/) by Nacho Frutos
Taken from: 圖像0362
(http://www.flickr.com/photos/26051511@N08/3444556930/)
by Cheunyiyn
Taken from: IMG_0242
(http://www.flickr.com/photos/98234419@N00/3444556920/) by Niblit_lvl50
Taken from: DMC-FS3 1/100 s de f/3,0 ISO 200
Objetivo Velocidad 1/100 s no flash
(http://www.flickr.com/photos/jpascualar/3444556912/) by Juan Ignacio Pascual
Taken from: D1010001 (http://www.flickr.com/photos/ihateselfish/3444556896/) by ihateselfish
Taken from: DSC00729
(http://www.flickr.com/photos/burmey/3443740213/
/) by Yumi Burmeister
# Taken from: R0010931
(http://www.flickr.com/photos/10fu/3443740185/)
by nacazy10fu
The nitty-gritty of backlinks

- Hyperlinks
- Search engine indexing
- Opacity
Technical problems with backlinks

- Multiple URLs for the same content
- Dynamic content
- Database searches
- Calendars
- 404 errors (no such webpage)
- Irrelevant parameters following '?' in URL
(and software)

- Bytes
- Lines of code
- Source files
- Package/project
(and images)

- Bytes
- Pixels
- Files
The fundamental problem

- Small content vs. large content
- The encyclopedia analogy
- The different types of bias
Alternatives

- The impossible ideal: a domain expert decides
- The page has real content, not found elsewhere
- The page has *some* content
- Impromptu licences
- Grouping
Grouping

- Domains
- Connected components
- Domain-isolated connected components
Comparison of methods, 2007

Number of URLs in group vs. rank.

- Blue triangles: domains
- Red triangles: subwebs
- Yellow triangles: both
# Largest .edu.au domains, 2007

<table>
<thead>
<tr>
<th>Rank</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>cddu.cqu.edu.au</td>
</tr>
<tr>
<td>2</td>
<td>parsa.rsise.anu.edu.au</td>
</tr>
<tr>
<td>3</td>
<td>mashedlc.edu.au</td>
</tr>
<tr>
<td>4</td>
<td>apsa.anu.edu.au</td>
</tr>
<tr>
<td>5</td>
<td>griffith.edu.au</td>
</tr>
<tr>
<td>6</td>
<td>community.mashedlc.edu.au</td>
</tr>
<tr>
<td>7</td>
<td>epinet.anu.edu.au</td>
</tr>
<tr>
<td>8</td>
<td>decide.it.uts.edu.au</td>
</tr>
<tr>
<td>9</td>
<td>images.swinburne.edu.au</td>
</tr>
<tr>
<td>10</td>
<td>myresearchspace.grs.uwa.edu.au</td>
</tr>
</tbody>
</table>
The Copyright Act

- There is copyright in a work if:
  - The author is Australian; or
  - The first publication of the work was in Australia
- Not a solution, but a paradigm with which to evaluate a solution
The Unit of the Commons

Ben Bildstein
ben.bildstein@student.unsw.edu.au
cyberlawcentre.org/hoc
The Unit of the Commons

- A discussion of metrics for online commons quantification

Not worrying about definitions, because that's covered elsewhere in the conference
Not getting too technical
Overview: This talk is mostly a warning on how to view quantitative data, and partly a view of how we might make progress forward in this area.
Why quantify?

- Justifying commons theory as a field of research
- Informing commons theory
  - Regulation of public goods
  - The commons in economics
- Informing licensing organisations
- Informing law and policy makers

Current commons theory lacks quantitative data, and instead implicitly assumes that our general 'feel' for the state of the commons is accurate. Creative Commons' non-attribution licences were retired based on the realisation that they were not being used. Informing law and policy makers probably hasn't happened yet, but hopefully will.
Media

- Text
- Images
- Sounds
- Video
- Software
Overview of current quantification results

- Creative Commons: CC wiki says '130 million total works'
- FOSS: SourceForge has 'more than 230,000 software projects'
- GFDL: Wikipedia has 2.83 million articles in English
- Photos: 102,868,725 photos

http://wiki.creativecommons.org/License_statistics
http://apps.sourceforge.net/trac/sourceforge/wiki/What%20is%20SourceForge.net?
http://wikipedia.org/
http://www.flickr.com/creativecommons/

Most of those SourceForge projects are half-baked, dead projects. And I can say that, because one of them is mine.

That's an incredible number of licensed photos. And they're not bad quality. Here's a completely random selection of 10 photos from Flickr.
Taken from: 2005 - Dormida
(http://www.flickr.com/photos/nfjoral/3444556932/)
by Nacho Frutos
Taken from: 圖像0362
(http://www.flickr.com/photos/26051511@N08/3444556930/)
by Cheuniyin
Taken from: IMG_0242
(http://www.flickr.com/photos/98234419@N00/3444556920/) by Niblit_lvl50
Taken from: DMC-FS3 1/100 s de f/3,0 ISO 200
Objetivo Velocidad 1/100 s no flash
(http://www.flickr.com/photos/jpascualar/3444556912/) by Juan Ignacio Pascual
Taken from: D1010001 (http://www.flickr.com/photos/ihateselfish/3444556896/) by ihateselfish
Taken from: DSCN3529
(http://www.flickr.com/photos/26018555@N04/3444556878/) by vipprettyhotmail_info
Taken from: DSC00729
(http://www.flickr.com/photos/burmey/3443740213/
/) by Yumi Burmeister
# Taken from: R0010931
(http://www.flickr.com/photos/10fu/3443740185/)
by nacazy10fu
Taken from: DSC01329 (http://www.flickr.com/photos/31233069@N05/3443740157/) by AppleParis
Going back to looking at Creative Commons' "130 million" figure now...
The nitty-gritty of backlinks

- Hyperlinks
- Search engine indexing
- Opacity

What Creative Commons is trying to do is to count hyperlinks to their licences. What they're actually doing is consulting an oracle: an opaque search engine, the workings of which we can't know, and which can change arbitrarily, without notice, and even without us knowing
Nowhere that I'm aware of are these temporal inconsistencies explain.
The best way to understand this graph is to recognise that any vertical slice represents Creative Commons best guess, at that time, or the real number of links to licences.

http://a6.creativecommons.org/~paulproteus/charts/2009-02-15/all/
Technical problems with backlinks

- Multiple URLs for the same content
- Dynamic content
- Database searches
- Calendars
- 404 errors (no such webpage)
- Irrelevant parameters following '?' in URL

E.g. Of multiple URLs: en.wikipedia.org/wiki/USA and en.wikipedia.org/wiki/United_States
Database search means a page of search results, of which there are clearly infinitely many possible sets of results, each of which might look like a licensed document
My SourceForge project is 1 package, with 8 files, 4853 lines of code, 160,000 bytes of source code.
(and images)

- Bytes
- Pixels
- Files
The fundamental problem

- Small content vs. large content
- The encyclopedia analogy
- The different types of bias

The encyclopedia analogy runs like this: imagine you wanted to collect all of the content on the web and put it in a book, like a huge encyclopedia. Then consider which parts of the web would go in, which wouldn't on account of being covered elsewhere, and which wouldn't on account of really not having any content.

URLs bias results towards pages that are accessible via multiple URLs
Web pages (where possible): bias towards document that are in many pieces
Domain names: bias towards domains with less content
Alternatives

- The impossible ideal: a domain expert decides
- The page has real content, not found elsewhere
- The page has *some* content
- Impromptu licences
- Grouping

These alternatives are either hard, or they are hacks. For perspective, the default methodology belongs in the 'hack' category.
Domain expert: hard (impossible)
Real content: hard, though hashing based on chunks of text like paragraphs may work
Impromptu licensed (a little hard, a little hack)
Grouping

- Domains
- Connected components
- Domain-isolated connected components

"Domains": group pages on the same domain into the same document
"Connected components": group pages that have links to each other into the same 'document'
"Domain-isolated connected components": as above, but cut groups into multiple groups where they exist on multiple domains.
Shows size of 'groups' from .edu.au domains in 2007. Data provided by National Library of Australia.
## Largest .edu.au domains, 2007

- 52590  cddu.cqu.edu.au
- 25122  parsa.rsise.anu.edu.au
- 24993  mashedlc.edu.au
- 21487  apsa.anu.edu.au
- 16747  griffith.edu.au
- 13748  community.mashedlc.edu.au
-  8606  epinet.anu.edu.au
-  1587  decide.it.uts.edu.au
-  1060  images.swinburne.edu.au
-   523  myresearchspace.grs.uwa.edu.au

Clearly none of these sites represent 10's of thousands of instances of someone writing a web page, and uploading it to their server. griffith.edu.au is an example of one with no content, just dynamic content, with multiple URLs for the same page.
The Copyright Act

- There is copyright in a work if:
  - The author is Australian; or
  - The first publication of the work was in Australia

- Not a solution, but a paradigm with which to evaluate a solution

Suggest the idea of 'published at the same time' as a way of grouping web pages into a single 'document'.
The Unit of the Commons

Ben Bildstein
ben.bildstein@student.unsw.edu.au
cyberlawcentre.org/hoc