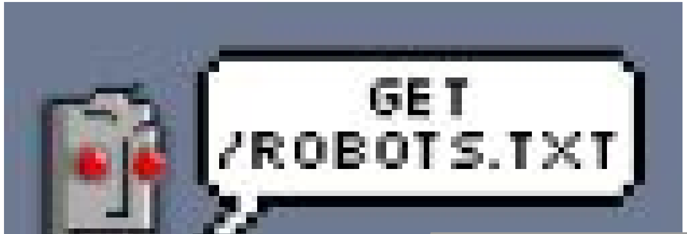
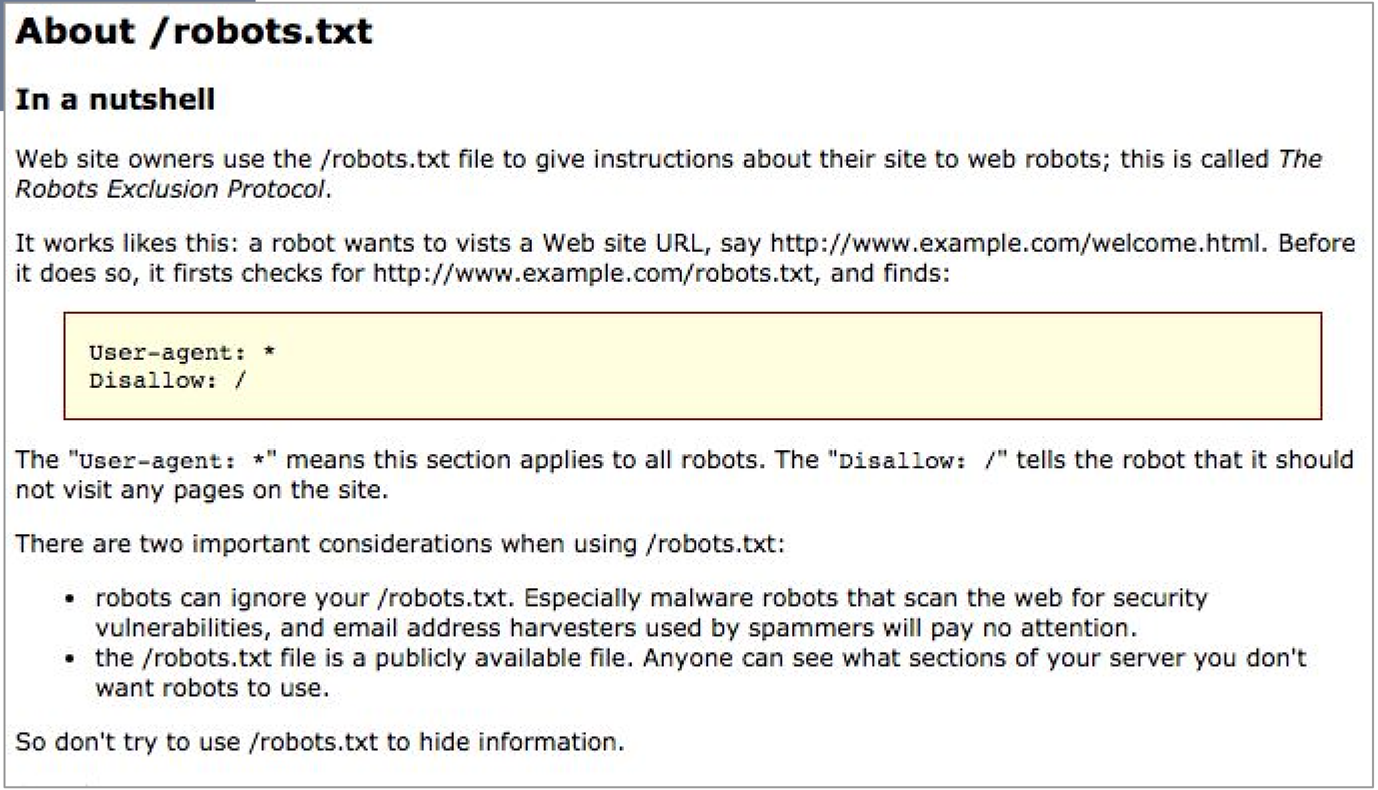
Robots in the Archives

Toke Eskildsen, Emily Maemura, Yves Maurer, Graham Seaman

Archives Unleashed | British Library | June 14, 2017



http://www.robotstxt.org/



About /robots.txt

In a nutshell

Web site owners use the /robots.txt file to give instructions about their site to web robots; this is called *The Robots Exclusion Protocol*.

It works likes this: a robot wants to vists a Web site URL, say http://www.example.com/welcome.html. Before it does so, it firsts checks for http://www.example.com/robots.txt, and finds:

```
User-agent: *
Disallow: /
```

The "`User-agent: *`" means this section applies to all robots. The "`Disallow: /`" tells the robot that it should not visit any pages on the site.

There are two important considerations when using /robots.txt:

- robots can ignore your /robots.txt. Especially malware robots that scan the web for security vulnerabilities, and email address harvesters used by spammers will pay no attention.
- the /robots.txt file is a publicly available file. Anyone can see what sections of your server you don't want robots to use.

So don't try to use /robots.txt to hide information.

The Question

What do we miss when we respect robots.txt exclusions?

Method:

Extract all robots.txt from the WARC collection

Apply it retroactively (nodejs robots-parser)

Idea : overlay the blocked resources on the link graph

How many robots.txt in a collection?

The National Archives

2010 Elections Collection, using sample from Department of Energy & Climate Change

478 domains *with* `robots.txt`

426 domains *without* `robots.txt`

http://webarchive.nationalarchives.gov.uk/20101213181030/http://www.decc.gov.uk

User-agents

Count	User-agent specified
857	User-agent: *
27	googlebot
11	msnbot
5	baiduspider
4	yahoo
4	ia_archiver

but are these outdated files?

```
. . .
Lycos_Spider_(T-Rex)
Jeeves v0.05alpha
```

www.bia.homeoffice.gov.uk
www.ind.homeoffice.gov.uk
www.ukba.homeoffice.gov.uk
www.audit-commission.gov.uk

Crawl-delay Directive

Count	Crawl-delay specified (in seconds)
35	`Crawl-delay: 10`
4	`Crawl-delay: 30`
4	`Crawl-delay: 60`
8	`Crawl-delay: 120`
2	`Crawl-delay: 300`
7	`Crawl-delay: 3600`

Focusing on links in decc.gov.uk subcollection

- Total links: 1,505,360
- Target outside of corpus: 90,087
- Unique targets: 20,506

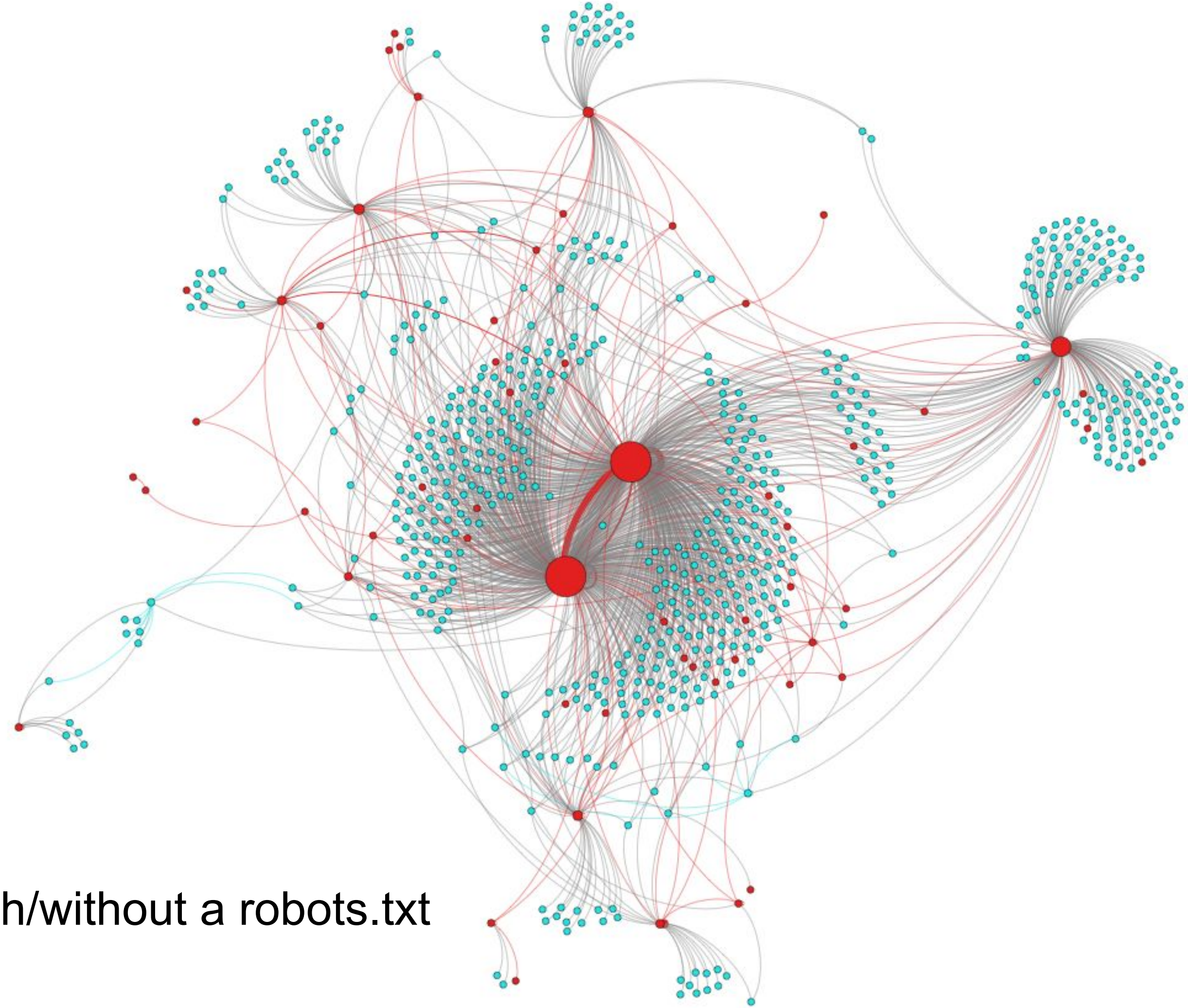
Retroactively Applying robots.txt to DECC

1 dimension: applying exclusions to individual URLs in WARCs

- 110,937 URLs Allowed
- 4,394 URLs Disallowed

2 dimensions: extracting links with warcbase

- Total links: 1,505,360
- # disallowed?



Domains with/without a robots.txt

Following links and obeying robots.txt

http://man270109a.decc.gov.uk/en/content/cms/what_we_do/change_energy/tackling_clima/int_climate/climate_news/prospectus_lau/prospectus_lau.aspx

http://www.decc.gov.uk/media/viewfile.ashx?filepath=internationalclimatechange/1_20100331081708_e_@@_beyondcopenhagen.pdf&filetype=4

http://man270109a.decc.gov.uk/en/content/cms/what_we_do/uk_supply/energy_mix/renewable/planning/plan_policy/north_ireland/north_ireland.aspxhttp://www.planningni.gov.uk/AreaPlans_Policy/PPS/pps1/PPS1.pdf

http://man270109a.decc.gov.uk/en/content/cms/what_we_do/uk_supply/energy_mix/renewable/planning/plan_policy/north_ireland/north_ireland.aspxhttp://www.planningni.gov.uk/AreaPlans_Policy/Supplementary/DCAN/dcan10/Dcan10.pdf

http://man270109a.decc.gov.uk/en/content/cms/what_we_do/uk_supply/energy_mix/renewable/policy/offshore/wind_leasing/wind_leasing.aspx

http://www.decc.gov.uk/media/viewfile.ashx?filepath=what%20we%20do/uk%20energy%20supply/energy%20mix/renewable%20energy/policy/offshore/wind_leasing/file51989.pdf&filetype=4

http://nora.nerc.ac.uk/ http://nora.nerc.ac.uk/cgi/search

http://nora.nerc.ac.uk/cgi/latest_tool?mode=bgs2010 http://nora.nerc.ac.uk/

http://nora.nerc.ac.uk/cgi/latest_tool?mode=bgs2010 http://nora.nerc.ac.uk/cgi/search

http://nora.nerc.ac.uk/nora/images/excel.gif http://nora.nerc.ac.uk/cgi/search

http://nora.nerc.ac.uk/nora/images/pdf.gif http://nora.nerc.ac.uk/cgi/search

http://nora.nerc.ac.uk/nora/images/ppt.gif http://nora.nerc.ac.uk/cgi/search

http://nora.nerc.ac.uk/nora/images/word.gif http://nora.nerc.ac.uk/cgi/search

http://nora.nerc.ac.uk/nora/images/zip.gif http://nora.nerc.ac.uk/cgi/search

http://www.bgs.ac.uk/mineralsuk/downloads/mpfcoal.pdf http://nora.nerc.ac.uk/cgi/latest_tool?mode=bgs2010

http://www.bgs.ac.uk/mineralsuk/downloads/mpfoilgas.pdf http://nora.nerc.ac.uk/cgi/latest_tool?mode=bgs2010

http://www.decc.gov.uk/en/content/cms/what_we_do/change_energy/tackling_clima/int_climate/climate_news/prospectus_lau/prospectus_lau.aspx

http://www.decc.gov.uk/media/viewfile.ashx?filepath=internationalclimatechange/1_20100331081708_e_@@_beyondcopenhagen.pdf&filetype=4

http://www.decc.gov.uk/en/content/cms/what_we_do/uk_supply/energy_mix/renewable/planning/plan_policy/north_ireland/north_ireland.aspx http://www.planningni.gov.uk/AreaPlans_Policy/PPS/pps1/PPS1.pdf

http://www.decc.gov.uk/en/content/cms/what_we_do/uk_supply/energy_mix/renewable/planning/plan_policy/north_ireland/north_ireland.aspx http://www.planningni.gov.uk/AreaPlans_Policy/Supplementary/DCAN/dcan10/Dcan10.pdf

http://www.decc.gov.uk/en/content/cms/what_we_do/uk_supply/energy_mix/renewable/policy/offshore/wind_leasing/wind_leasing.aspx

http://www.decc.gov.uk/media/viewfile.ashx?filepath=what%20we%20do/uk%20energy%20supply/energy%20mix/renewable%20energy/policy/offshore/wind_leasing/file51989.pdf&filetype=4

https://www.energynpsconsultation.decc.gov.uk/nuclear/managementdisposalwaste/annex/wastes/storage/ http://mrws.decc.gov.uk/media/viewfile.ashx?filepath=mrws/white-paper-final.pdf&filetype=4

https://www.energynpsconsultation.decc.gov.uk/nuclear/managementdisposalwaste/annex/wastes/suitable/ http://mrws.decc.gov.uk/media/viewfile.ashx?filepath=mrws/white-paper-final.pdf&filetype=4

https://www.og.decc.gov.uk/EIP/pages/onshore.htm

http://www.decc.gov.uk/media/viewfile.ashx?filepath=what%20we%20do/uk%20energy%20supply/development%20consents%20and%20planning%20reform/electricity/1_20091106164611_e_@@_ccrguidance.pdf&filetype=4

https://www.og.decc.gov.uk/information/fields.htm http://www.decc.gov.uk/media/viewfile.ashx?filepath=statistics/source/oil/dukesf_1.xls&filetype=4

https://www.og.decc.gov.uk/information/fields.htm http://www.decc.gov.uk/media/viewfile.ashx?filepath=statistics/source/oil/dukesf_2.xls&filetype=4

https://www.og.decc.gov.uk/regulation/pons/pon_09b.htm http://www.og.decc.gov.uk/pls/wons/wonsw001.loginform

warcbase link extractor != harvester link extractor

Conclusion

- Impact of robots.txt is minimal for this collection
- Our method can be applied to other collections and extended to further the discussion on ignore robots.txt -- how to determine if, how, or when to ignore it

Extra Slides

Estimating the age of `robots.txt` files

Older User-agents listed:

- Lycos_Spider_(T-Rex)
- Jeeves v0.05alpha (PERL, LWP, lglb@doc.ic.ac.uk)
- Architext (Excite)
- asterias (AOL)

list of user agents: www.user-agents.org

Minimum viable product

Applying all URLs to the domain-specific robots.txt

	Allowed	Disallowed
Python (robotexclusionrulesparser)	111,002	4,329
Nodejs (robotsparser)	110,937	4,394

Minimum viable product

Applying all URLs to the domain-specific robots.txt

	Allowed	Disallowed
Python (robotexclusionrulesparser)	111,002	4,329
Nodejs (robotsparser)	110,937	4,394

But…

http://www.decc.gov.uk/media/viewfile.ashx?filepath=what%20we%20do/uk%20energy%20supply/energy%20mix/nuclear/nonproliferation/chemicalbiologicalweapons/cwc_uk_auth/file37019.pdf&filetype=4

has 4,234 variants, which leaves only 160 other URLs

Misc. observations

- Warcbase link extractor != harvester link extractor
  - 1 case ⇒ 4,234 instances
- Near-duplicate content (crawler-traps) contaminate statistics
- 1.5M links ⇒ 20K unique targets

For this corpus, the impact was minimal.

Our method is easily applicable for other corpora - we should do that.

The problem

```
//<![CDATA[
var flashvars = {};
var params = { menu:
'false',allowFlashAutoInstall:'false',Flashvars:'&HostURL= http://www.decc.gov.
uk%2fmedia%2fviewfile.ashx%3ffilepath%3dimagearray.xml%26Component%3dHomeFlash
Banner&autoPlay=true&Delay=10000 ',
allowScriptAccess:'sameDomain', movie:'/en/flash/decc.swf',
salign:'lt',quality:'high',scale:'noscale' };
var attributes = {id: 'mymovie','align':'top'};
swfobject.embedSWF('/Media/ViewFile.ashx?FileType=7&Component=HomeFlashBanner&
FilePath=ImageRotator.swf', 'FlashMarker', '580', '209', '6', false,
flashvars, params, attributes);
//]]>
```