On the Persistence of Persistent Identifiers of the Scholarly Web

Martin Klein
Los Alamos National Laboratory
martinklein0815@gmail.com
@mart1nkle1n
For more background, details, results


On the Persistence of Persistent Identifiers of the Scholarly Web
@mart1nkle1n
CNI Spring Virtual Meeting 2020
On the Persistence of Persistent Identifiers of the Scholarly Web
@mart1nkle1n
CNI Spring Virtual Meeting 2020
On the Persistence of Persistent Identifiers of the Scholarly Web
@mart1nkle1n
CNI Spring Virtual Meeting 2020
On the Persistence of Persistent Identifiers of the Scholarly Web
@mart1nkle1n
CNI Spring Virtual Meeting 2020
On the Persistence of Persistent Identifiers of the Scholarly Web
@mart1nkle1n
CNI Spring Virtual Meeting 2020
On the Persistence of Persistent Identifiers of the Scholarly Web
@mart1nkle1n
CNI Spring Virtual Meeting 2020
On the Persistence of Persistent Identifiers of the Scholarly Web
@mart1nkle1n
CNI Spring Virtual Meeting 2020
On the Persistence of Persistent Identifiers of the Scholarly Web
@mart1nkle1n
CNI Spring Virtual Meeting 2020
Regardless of location, phone used ...

... when calling a well-known number that uniquely identifies a(n) (emergency) resource ...

... would you not expect to get the same response?

Do you trust in the persistence of that number (and the response)?
No more scary emergency scenarios!

- Phones == web clients
- Locations == network environments
- 911 calls == HTTP requests against DOIs
- Regardless of the web client and network location, would you not expect the same response from a web server when requesting the same DOI?
Idea...

- Comparative study investigating scholarly publishers’ responses
  - To common HTTP requests
  - Against DOIs

- Using different web clients and request methods, resembling
  - Machines "browsing", crawling
  - Humans browsing

- From network environments with different subscriptions/licenses
  - Amazon Web Service EC2 instance
  - LANL internal
How does this work?

10.1007/978-3-540-87599-4_38
How does this (not) work?

10.1007/978-3-540-87599-4_38
How does this work via HTTP?

https://doi.org/10.1007/978-3-540-87599-4_38
What do you see?

https://doi.org/10.1007/978-3-540-87599-4_38

Revisiting Lexical Signatures to (Re-)Discover Web Pages

Martin Klein, Michael L. Nelson

Abstract

A lexical signature (LS) is a small set of terms derived from a document that capture the "aboutness" of that document. A LS generated from a web page can be used to discover that page at a different URL as well as to find relevant pages in the Internet. From a set of randomly selected URLs we took all their copies from the Internet Archive between 1996 and 2007 and generated their LSs. We conducted an overlap analysis of terms in all LSs and found only small overlaps in the early years (1996 - 2000) but increasing numbers in the more recent past (from 2003 on). We measured the performance of all LSs in dependence of the number of terms they
How does this work via HTTP?

https://doi.org/10.1007/978-3-540-87599-4_38
How does this work via HTTP?

https://doi.org/10.1007/978-3-540-87599-4_38

(HTTP redirect)

http://link.springer.com/10.1007/978-3-540-87599-4_38
How does this work via HTTP?

https://doi.org/10.1007/978-3-540-87599-4_38

(HTML redirect)

http://link.springer.com/10.1007/978-3-540-87599-4_38

(HTML redirect)

https://link.springer.com/10.1007/978-3-540-87599-4_38
How does this work via HTTP?

https://doi.org/10.1007/978-3-540-87599-4_38

(HTTPS redirect)

http://link.springer.com/10.1007/978-3-540-87599-4_38

(HTTPS redirect)

https://link.springer.com/10.1007/978-3-540-87599-4_38

(HTTPS redirect)

https://link.springer.com/chapter/10.1007%2F978-3-540-87599-4_38
DOI dataset

• Gathering a representative sample is not trivial!

• Internet Archive conducts crawls of the scholarly domain
  • June 2018: 93 million DOIs
  • Obtained WARC files and extracted DOI redirect chain

• Investigate publisher distribution
  • Final link of redirect chain and extract host e.g.:
    https://link.springer.com/chapter/10.1007%2F978-3-540-87599-4_38
    Domain: springer.com

• Randomly pick 100 DOIs from the 100 most frequent domains
  • 10,000 DOIs
Web clients and HTTP requests 1/4

- **HEAD request**
  - Server responds with response headers
  - *but no* response body
  - Client: cURL

```
curl -IL http://doi.org/10.1016/j.wocn.2010.05.003

HTTP/1.1 302
Date: Sat, 16 Mar 2019 23:20:13 GMT
Content-type: text/html; charset=utf-8
Content-Length: 209
Location: https://linkinghub.elsevier.com/retrieve/pii/S0095...

HTTP/1.1 200
Date: Sat, 16 Mar 2019 23:20:14 GMT
Content-Type: text/html; charset=UTF-8
Content-Language: en-US
```
Web clients and HTTP requests 1/4

- **HEAD request**
  - Server responds with response headers
  - *but no* response body
  - Client: cURL

```bash
```

HTTP/1.1 302
Date: Sat, 16 Mar 2019 23:20:13 GMT
Content-Type: text/html; charset=utf-8
Content-Length: 209
Location: https://linkinghub.elsevier.com/retrieve/pii/S027853191930050X

HTTP/1.1 200
Date: Sat, 16 Mar 2019 23:20:14 GMT
Content-Type: text/html; charset=UTF-8
Content-Language: en-US
**Web clients and HTTP requests 2/4**

- **GET request**
  - Server responds with response headers
  - *and* response body
  - Client: `cURL`

```bash
curl -iL http://doi.org/10.1016/j.wocn.2010.05.003
```

HTTP/1.1 302
Date: Sat, 16 Mar 2019 23:20:13 GMT
Content-Type: text/html; charset=utf-8
Content-Length: 209
Location: https://linkinghub.elsevier.com/retrieve/pii/S0095...

HTTP/1.1 200
Date: Sat, 16 Mar 2019 23:20:14 GMT
Content-Type: text/html; charset=UTF-8
Content-Language: en-US

```html
<!DOCTYPE HTML PUBLIC ... >
<html>
<head>
    ..... 
</head>
<body>
    ..... 
</body>
</html>
```
Web clients and HTTP requests 2/4

- **GET** request
  - Server responds with response headers
  - *and* response body
  - Client: cURL

```bash
curl --loadout http://doi.org/10.1016/j.wocn.2010.05.001
```

```
HTTP/1.1 302
Date: Sat, 16 Mar 2019 23:20:13 GMT
Content-Type: text/html; charset=utf-8
Content-Length: 209
Location: https://linkinghub.elsevier.com/retrieve/pii/S0890812310000154

HTTP/1.1 200
Date: Sat, 16 Mar 2019 23:20:14 GMT
Content-Type: text/html; charset=UTF-8
Content-Language: en-US

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<html>
<head>
  ...
</head>
<body>
  ...
</body>
</html>
```
Web clients and HTTP requests 3/4

- GET+
  - GET request with request headers
    - User Agent (desktop Chrome browser)
    - Specified connection timeout
    - Specified maximum number of redirects
    - Cookies accepted and stored
    - Insecure connections allowed
  - Client: cURL
Web clients and HTTP requests 3/4

- **GET+**
  - GET request with request headers
    - User Agent (desktop Chrome browser)
    - Specified connection timeout
    - Specified maximum number of redirects
    - Cookies accepted and stored
    - Insecure connections allowed
  - Client: cURL
Web clients and HTTP requests 4/4

- **Chrome:**
  - GET request via Selenium Webdriver controlled browser
  - Client: Chrome

---

On the Persistence of Persistent Identifiers of the Scholarly Web
@mart1nkle1n
CNI Spring Virtual Meeting 2020
Web clients and HTTP requests 4/4

- **Chrome:**
  - GET request via Selenium Webdriver controlled browser
  - Client: Chrome

A lexical signature (LS) is a small set of terms derived from a document and reflects its "aboutness". A LS generated from a web page at a different URL, as well as to find relevant pages in the Internet Archive, we took all their copies from the Internet Archive generated their LSs. We conducted an overlap analysis of terms in these LSs, which overlaps in the early years (1996 – 2000) but increasing number on (2003 on). We measured the performance of all LSs in dependence of overlap and found that LSs created more recently performed better.
Regarding response headers, RFC 7231 states:
(highlights mine)

“The server **SHOULD** send **the same header fields** in response to a **HEAD** request as it would have sent if the request had been a **GET**...”.

HTTP response codes

- **2xx**
  - Success

- **3xx**
  - Redirection

- **4xx**
  - Client error

- **5xx**
  - Server error
Response codes of last link in redirect chain by DOI

On the Persistence of Persistent Identifiers of the Scholarly Web
@mart1nkle1n
CNI Spring Virtual Meeting 2020
Response codes of last link in redirect chain by DOI

• < 50% successful requests across all methods

48.3%
Response codes of last link in redirect chain by DOI

- < 50% successful requests across all methods
- > 40% 300-level responses w/ GET
Response codes of last link in redirect chain by DOI

- < 50% successful requests across all methods
- > 40% 300-level responses w/ GET
  - 25% of them 200-level w/ HEAD/Chrome
Response codes of last link in redirect chain by DOI

- < 50% successful requests across all methods
- > 40% 300-level responses w/ GET
  - 25% of them 200-level w/ HEAD/Chrome
- 13% 400-level responses w/ HEAD
Response codes of last link in redirect chain by DOI

- < 50% successful requests across all methods
- > 40% 300-level responses w/ GET
  - 25% of them 200-level w/ HEAD/Chrome
- 13% 400-level responses w/ HEAD
  - 25% of them w/ 200-level response w/ any other method
Response code comparison external vs internal network

![Diagram comparing response codes for HEAD, GET, GET+, and Chrome with external and internal networks.]

- **External Network**
  - 2xx: 48.3%
  - 3xx: 6.9%
  - 4xx: 12.3%
  - 5xx: 1.7%
  - 5xx: 2.5%

- **Internal Network**
  - 2xx: 66.9%
  - 3xx: 3.1%
  - 4xx: 5.3%
  - 5xx: 1.8%
  - 5xx: 2.5%

On the Persistence of Persistent Identifiers of the Scholarly Web

@mart1nkle1n

CNI Spring Virtual Meeting 2020
Response code comparison OA vs non-OA

On the Persistence of Persistent Identifiers of the Scholarly Web
@mart1nkle1n
CNI Spring Virtual Meeting 2020
Response code comparison SUB vs non-SUB

SUB

non-SUB

1,266 DOIs

8,734 DOIs

10,000 DOIs
Take-aways

• Frequently, scholarly publishers respond inconsistently to different requests against the same DOI, depending on:
  • HTTP client, request method, network environment

• Implications for (perceived) persistence of DOIs?
  • Inconsistent DOI resolution does not build trust in DOIs
  • Lack of adherence to standards does not build trust

• More work needed but initial findings seem to indicate:
  • OA DOIs more consistent than non-OA DOIs
  • DOIs for subscribed & licensed content show more consistency

• Implications for archival efforts?
  • Test different combinations of clients/request methods/networks
  • Pretend to be as human as possible
Photo credits

- https://unsplash.com/photos/SBiVq9eWEtQ
- https://unsplash.com/photos/9e9PD9blAto
- https://unsplash.com/photos/97HfVpyNR1M
- https://unsplash.com/photos/UYwjKbrwUos
- https://unsplash.com/photos/Se7vVKzYxTI
- https://unsplash.com/photos/_geAgtjqLzY
- https://unsplash.com/photos/r–enAOPw8Rs
- https://unsplash.com/photos/A4qmsfG6ywM
- https://unsplash.com/photos/eWqOgJ–Ifl
- https://unsplash.com/photos/goholCAVTRs
- https://unsplash.com/photos/vpXbwh6Qk9U
- https://unsplash.com/photos/K21Dn4OVxNw
- https://unsplash.com/photos/HzOclMmYryc
- https://unsplash.com/photos/OW5KP_Pj85Q
On the Persistence of Persistent Identifiers of the Scholarly Web

Martin Klein
Los Alamos National Laboratory
martinklein0815@gmail.com
@mart1nkle1n

Thank you & stay safe!