What does it mean when your URIs are redirected? Examining identity and redirection in the LOD cloud

Idries Nasim, Shuai Wang, Joe Raad, Peter Bloem, Frank van Harmelen

VU Amsterdam
Paris-Saclay University

Outline

Introduction

Preliminaries

Related Research

Research Questions

Data Preparation

Implicit Semantics of Redirection - RQ1

Analyzing the Redirection Graph - RQ2

Discussion and Conclusion

Introduction

- The semantic web is a decentralised world-wide information space
- Resources are identified by Uniform Resource Identifiers (URI).
- Outdated resources are typically redirect a new location
- We study how redirection indicate evolution of entities in the LOD cloud
- From non-information resource to information resource

For example, https://www.worldcat.org/oclc/67950327 redirects to https://www.worldcat.org/title/pro-patria/oclc/67950327

Preliminaries

Identity links (black edges)

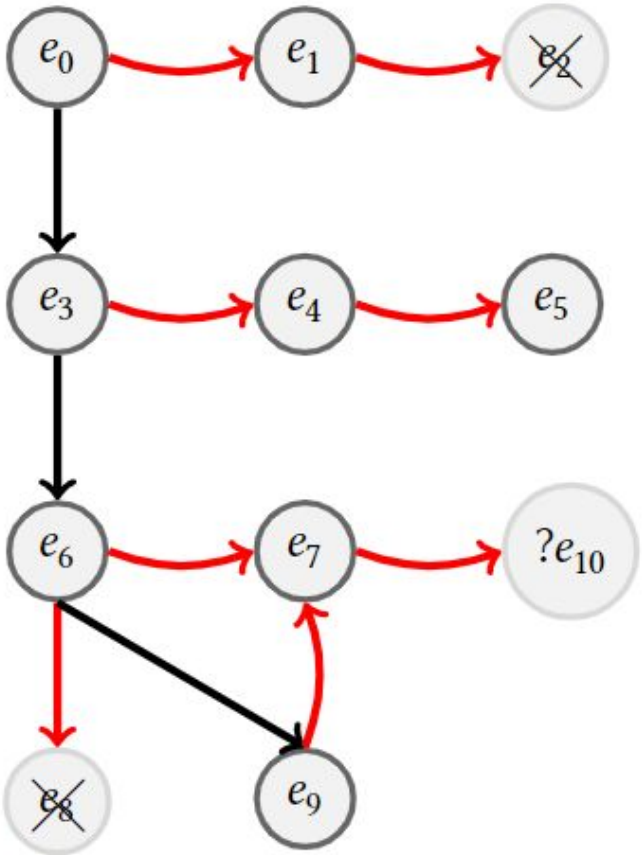
Identity graphs (e.g. owl:sameAs)

Redirection (HTTP 3XX/Hash convention, red edges)

Redirection graph

Issues:

- Not found (HTTP 400+)
- Timeout
- Redirected until not found
- Error
- …



Related Research

The identity crisis + evolution was studied by Halpin, et al.

- propose to study how an HTTP resource responds to a GET request
- Ontology
- No quantitative analysis

19.4% entities no longer exist on the web after 2 years, found by De Melo.

- BTC 2011 DBpedia
- Escaped titles
- Wikipedia is a living resource

Regino et al. studied broken links

- Refer to different entities after evolution
- Wikidata, GeoNames, BabelNet
- Tracking every version of every entity in every dataset is impossible

This morning: Pieter Colpaert’s work on sometimes available APIs.

Research Questions

RQ1: Can we approximate the implicit semantics of the redirection?

RQ2: What are the properties and structure of redirection graphs?

Data Preparation

Sameas.cc is the identity graph corresponding to the 2015 LOD Laundromat datasets (650K)

- 558.9M owl:sameAs links
- 179.7M entities

Sampling (to understand if size of CCs has to do with redirection)

1. Uniform sampling (100K)
2. Connected components (CCs) of 2 entities
3. Connected components (CCs) of 3-10 entities
4. Connected components (CCs) of 10+ entities

Data Preparation: redirection graph

Record the results of HTTP GET request

- HTTP 200: OK
- Timeout: TO
- Not Found: NF
- Error: ER

Redirection also includes:

  - ➢ 300 (redirection with multiple choice)
  - ➢ 301 (moved permanently)
  - ➢ 307 (temporal redirect)
  - ➢ 308 (permanent redirect), etc.
- Redirect until timeout (RUT)
- Redirected until error (RUE)
- Redirected until found (RUF)
- Redirected until not found (RUNF)

Timeout

- connection timeout to 0.01s and read timeout 0.05s
- then 0.5 and 2.5 seconds
- Then 5 and 25 seconds

Implicit Semantics

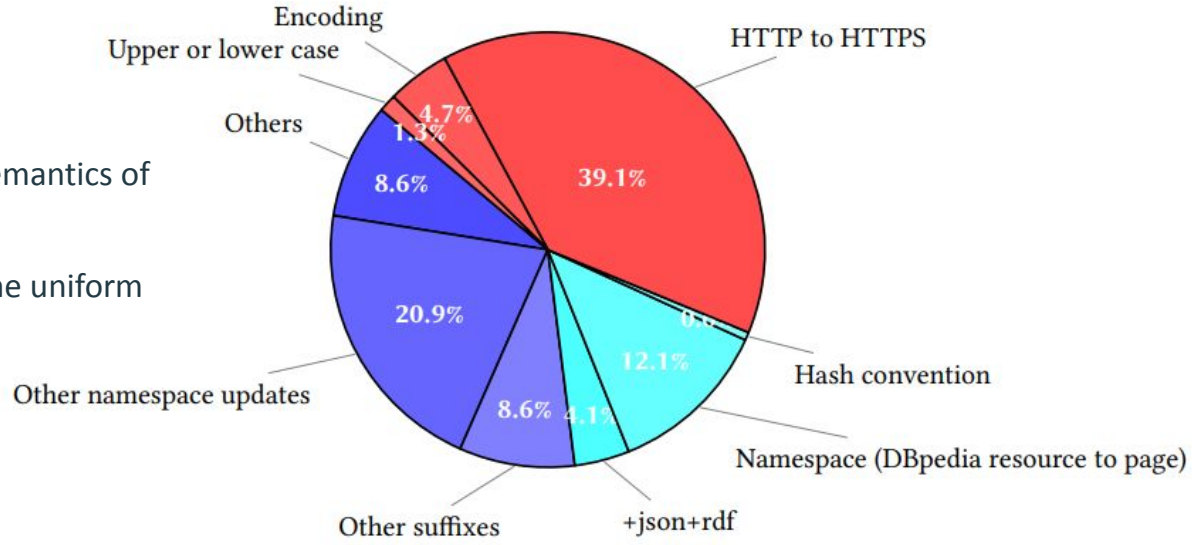
RQ1: Can we approximate the implicit semantics of the redirection?

4,000 manually annotated edges from the uniform sampling

39.1% about HTTP to HTTPS

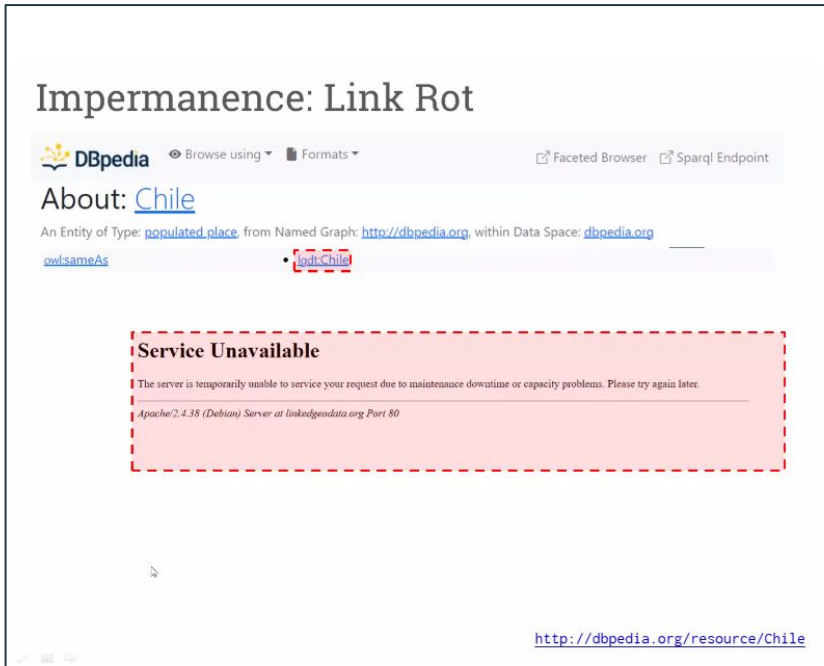
Encoding / upper lower case

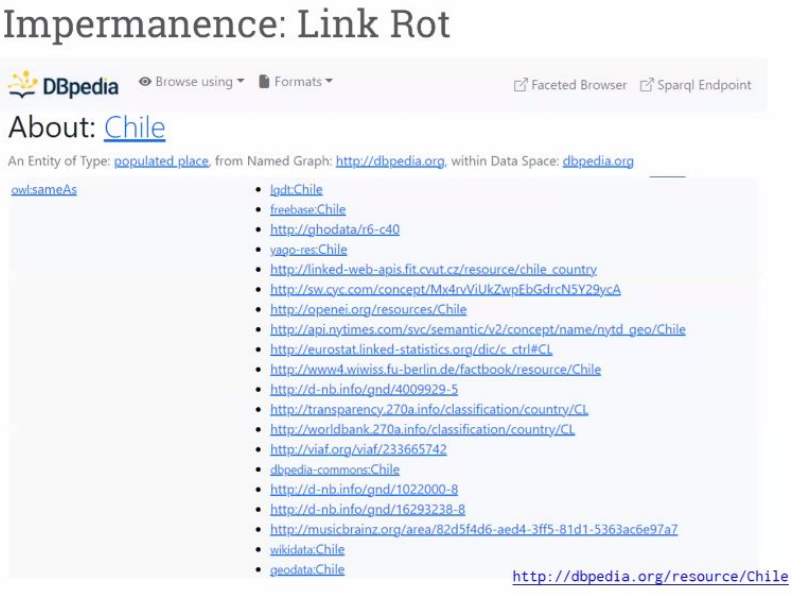
Hash convention is not as common



Between 45.1% and 83.2% of redirection links can be taken as identity links

Recall Aidan's presentation just now





Implicit Semantics

1.1% returns HTTP 200
Only 33.7% valid URIs (some information after dereferencing)

Table 1: Behavior of HTTP GET request of entities

Graph	RUF	OK	Valid[1]	ER	TO	RUT	RUNF	RUE	NF	Invalid[2]
R^{U}	32.6%	1.1%	33.7%	23.9%	8.2%	8.1%	12.8%	0.01%	13.3%	66.3%
$R^{CC(2)}$	37.1%	0.7%	37.8%	39.5%	12.3%	0.9%	5.5%	0.0%	4.0%	62.2%
$R^{CC(3-10)}$	30.4%	0.3%	30.7%	43.4%	5.8%	0.9%	5.8%	5.0%	8.4%	69.3%
$R^{CC(>10)}$	26.0%	0.8%	26.8%	26.5%	23.2%	2.3%	10.1%	0.1%	11.0%	73.2%

[1] The valid entities include RUF (redirected until found), OK (found with HTTP 200)
[2] The rest are invalid entities, including ER (error), TO (timeout), RUT (redirected until timeout), RUNF (redirected until not found), RUE (redirected until error), and NF (not found).

Implicit Semantics

Indicates that the greater the size of CCs increase, fewer valid URIs.
Not always positively or negatively correlating to the size of CC

Table 1: Behavior of HTTP GET request of entities

Graph	RUF	OK	Valid[1]	ER	TO	RUT	RUNF	RUE	NF	Invalid[2]
R^{U}	32.6%	1.1%	33.7%	23.9%	8.2%	8.1%	12.8%	0.01%	13.3%	66.3%
$R^{CC(2)}$	37.1%	0.7%	37.8%	39.5%	12.3%	0.9%	5.5%	0.0%	4.0%	62.2%
$R^{CC(3-10)}$	30.4%	0.3%	30.7%	43.4%	5.8%	0.9%	5.8%	5.0%	8.4%	69.3%
$R^{CC(>10)}$	26.0%	0.8%	26.8%	26.5%	23.2%	2.3%	10.1%	0.1%	11.0%	73.2%

[1] The valid entities include RUF (redirected until found), OK (found with HTTP 200)

[2] The rest are invalid entities, including ER (error), TO (timeout), RUT (redirected until timeout), RUNF (redirected until not found), RUE (redirected until error), and NF (not found).

Manual Examination

We extract 100 chains of redirection with at least 2 hops.

Mixed types of redirection

85% happens within a domain

- 28% wikidata
- 26% DBpedia (often resource to pages), also among the longest path
- Others: bibsonomy.org(5%) and viaf.org (1%)

[‘http://dbpedia.org/resource/Mirage_%28pop_group%29',
'http://dbpedia.org/resource/Mirage_(pop_group)',
'https://dbpedia.org/resource/Mirage_(pop_group)',
'http://dbpedia.org/page/Mirage_(pop_group)',
'https://dbpedia.org/page/Mirage_(pop_group)',
'http://dbpedia.org/resource/Mirage_(disambiguation)',
'https://dbpedia.org/resource/Mirage_(disambiguation)',
'http://dbpedia.org/page/Mirage_(disambiguation)',
‘https://dbpedia.org/page/Mirage_(disambiguation)']

Analyzing the Redirection Graphs

RQ2: What are the properties and structure of redirection graphs?

Redirection chains can be involve as many as 9 entities, only about DBpedia pages/resources

38-53% entities are redirected; very common!

$R^{CC(3\text{-}10)}$ has a cycle!

Large CCs could suggest poorly maintained URIs

Graph	#Entities	#Entities Redirected	#Nodes	#Edges	Avg #Hops	Max #Hops
R^{U}	100K	53,487 (53.49%)	169,021	116,031	1.71	8
$R^{CC(2)}$	20K	8,693 (43.46%)	30,091	21,602	1.64	8
$R^{CC(3-10)}$	20K	8,412 (42.06%)	29,697	21,490	-	-
$R^{CC(>10)}$	20K	7,704 (38.52%)	24,914	18,102	2.05	8

Discussion and Future work

Only studied entities in the identity graphs (very outdated)

A new identity graph (to submit to ESWC)

We did not study the entire identity graphs, but some samples of them

Although redirection relations cannot be used as identity relations, they can be useful.

Decentralized Identifiers (DIDs)

[learned this morning] Publish event streams, not datadumps; use materializable interfaces

Study the temporal aspect of URIs using the ontology proposed by Halpin, et al.

Contributions

- four redirection graphs corresponding to different sampling methods using the sameas.cc identity graph;
- 4,000 semi-automatically annotated edges (as pairs of URIs) in the uniformly sampled redirection graph;
- a qualitative study of the semantics of redirection in the identity graphs;
- a quantitative study of properties of the redirection graphs;

Data: https://doi.org/10.5281/zenodo.7225383
Code: https://github.com/shuaiwangvu/redirection

Contact:
shuai.wang@vu.nl