# HTTP vs. HTTPS in resource identification

In November 2018 at SWIB18 conference a breakout session "*Cool URIs might be insecure if they don't change*" took place. Participants wished to find a place to record the results and follow up on the discussion. This wiki page is meant to serve this purpose.

It is hosted by the German DINI-AG Competence Centre Interoperable Metadata (KIM). Further exchange of any parties interested in the topic could be initiated through comments on this page or through the mailing list of KIM's Working Group "Identifiers" (mailing list registration, maling list public archive).



# SWIB18 Breakout Session: Cool URIs might be insecure if they don't change

**2018-11-27. Bonn, Germany. Friedrich-Ebert-Stiftung.**

**Participants:** Lars G. Svensson (Deutsche Nationalbibliothek) (*Initiator*), Jana Hentschke (Deutsche Nationalbibliothek) (*Initiator*), Raphaëlle Lapôtre (Bibliothèque nationale de France), Pascal Christoph (Hochschulbibliothekszentrum NRW (hbz)), Michele Casalini (Casalini Libri), Carsten Klee (Staatsbibliothek zu Berlin - Preußischer Kulturbesitz), Tom Baker (Dublin Core Metadata Initiative), Alexander Jahnke (Niedersächsische Staats- und Universitätsbibliothek Göttingen), Martin Scholz (Friedrich-Alexander-Universität Erlangen-Nürnberg), Joachim Laczny (Staatsbibliothek zu Berlin – Preußischer Kulturbesitz)

**Teaser:**

*Many (most?) providers of linked data publish their resources using http URIs as identifiers. http, however, is a very insecure protocol and there is a movement towards making the web – and thus the Semantic Web – a more secure and trusted place by moving from http to https. For the cosmos of linked RDF data, the seemingly small addition of one character changes a lot: the URIs of a resource. And cool URIs don't change ...*

*Redirects from http to https URIs seem to solve the problem at first glance. But do they really? As long as an initial request for a resource is directed at its cool http-URI, there is unprotected exchange of data that could be intercepted (or even altered). On the other hand, changing http://example.org /resource123 to https://example.org/resource123 might cause some discomfort at the data consumer side as data stores need to be updated and queries amended.*

*How are "Cool URIs" to be weighed out against trustworthiness of data providers, privacy of users ... ?*

**Introductory slides:** Should the Semantic Web switch to HTTPS.pptx (by Jana)

**Discussion notes** (*arranged by subtopics by Jana, compiled from memory and collective etherpad notes*):

### Why can identification and data transfer not be considered two different things?

Statement #1 of the Linked Data Principles says "*Use URIs as names for things*" and statement #2 says "*Use http URIs to that people can look up those names*" (This includes https URIs, too!). This stresses that URIs are used for the purpose of identification as well as to initiate data transfer.

### Why do 301 redirects not solve the problem?

The problem is that http traffic can be intercepted, so redirects aren't a solution, since the initial http request can be intercepted and e.g. a 301 redirect can be replaced by a redirect to another resource

### Why do owl:sameAs relations not solve the problem?

On the data level they do solve the problem - when inferencing is applied (which is the idea about linking anyway - the linked data cosmos is principally able to handle different URIs for the same things).

On the protocol level, however, it doesn't help. When you have an http:xxx sameAs https:xxx you have to exploit the (insecure) http data first, before going to https. Any http transfer makes the data exchange insecure. So owl:sameAs relations between http and https URIs just blow up the data volume and doesn't solve the problem. According to Halpin there is no way in RDF to say that one URI is equivalent to another URI (owl:sameAs links individual /entities but not the identifiers).

### Is it really so important for our kind of data to make them interception-prove?

Noone has heard of a case of library data being manipulated in bad faith. However, this argument doesn't work out because we are talking about a matter of principle. Trustworthy data exchange is needed (no matter if libraries should be the first to demand it).

### Why are identifiers not independent from protocol?

The RDF specification says that identifiers are compared character bv character (as specified in RFC 3986). This means that for a generic RDF handler, http://example.org/foo and https://example.org/foo are different resources since a URI comparator would say that the fifth character in the string differs.

Linked Data, however, is about the web so we need web-actionable identifiers. Deploying something like a resolver in between breaks the web approach: you have to educate users to use the resolver instead (which btw. has other disadvantages - bad experience with resolvers with DOI because the resolvers behave so differently. BnF experience: It's not sufficient to set an ARK on everything)

Halpin thinks that it would be a logical step in the future for the RDF standard to declare https und http URIs equivalent. Recently there is much discussion on the semweb mailing list regarding features for RDF2 (incl. other topics such as literals as subjects, b-nodes as predicates etc.)

### Why not simply change to https? What happens, exactly, if URIs are not cool and change?

The problem is not your own data, but everyone else that uses it. Even with long-term 301 redirects from the http to the succeeding https URIs the clients /other systems still need to look up the data they already have to find out that there was a change and that two URIs refer to the same thing. The problem is when you have a mixture of legacy and new data, e.g. SPARQL queries will be much harder to write. It can be argued that this, today, affects a relatively small group which would need to adapt only this one time.

### Why not start assigning https identifiers to new resources while keeping http for the existing ones?

We could assign https to new data and keep http for "old" data. However, that would be awkwardly inconsistent and hard to explain to non-LD people, particularly if the URIs are based on patterns (e. g. "Use the prefix http://example.com/data/ and add the foo-number"). And still involves http traffic.

### How do search engines react?

Websites get downgraded by search engines if they don't serve https. Also, if you have mixed content, browsers will flag that.

### How to prevent mixed content warnings?

When http resources are linked from https resources, modern browsers will display "mixed content" warnings. Assuming there will always remain linked data providers that don't use https-URIs, mixed content warnings are likely to occur when data is being linked. This is the chicken or the egg problem. It can be considered another argument for doing nothing: if not everyone does it, the ecosystem will be http-based anyway. (This attitude, however, will not enable improvement ...)

### Can the problem with http identifiers be solved on the client side?

Why can't e.g. SPARQL clients secretly update to HTTPS before sending a query - just as modern browsers do.

Another idea: applying a checksum upon data transfer to make sure that data can not be corrupted? To prevent those checksums from being manipulated over http, the checksum would be the same in http and https, so you would be able to see that the checksum isn't the same over http and https.

Relying on technical advancement and client side is also what is recommended in a 2016 W3C Blog post: (summarised by Sandro Hawke) '*[...] keep writing “http:” and trust that the infrastructure will quietly switch over to TLS (https) whenever both client and server can handle it. Meanwhile, let’s try to get SemWeb software to be doing TLS+UIR+HSTS and be as secure as modern browsers.*'

The question is: why wait? Harry Halpin sees no reason to wait for RDF standard or client side changes because the (semantic) web is young enough to change to https.

### What about URIs of RDF Element Sets|Vocabularies

Does it make sense to change them to https as long as rdf, owl, skos, xsl ... don't change? Policy of these is to wait and hope that HSTS and UIR will still become good.

Another point is that the semantics of most of those vocabularies are hard-coded in the relevant tools (such as the OWL-API or Jena) so the identifiers are not dereferenced before they are acted on. That means that it would not really matter if the rdf namespace is identified by http://www.w3.org/1999/02/22-rdf-syntax-ns# or https://www.w3.org/1999/02/22-rdf-syntax-ns# since the semantics are known and the namespace identifiers are not dereferenced anyway. Terms, classes and properties from other vocabularies building on RDF, RDFS and OWL need to be dereferenced in order for applications to act on the properly, so here it does matter if the identifiers use http or https.

DCMI has had only one request for https so far and therefore currently doesn't see urgent need for action.

### Who does (or thinks) what?

Bibliothèque nationale de France has upgraded to https in the data on the web pages, but not in the SPARQL endpoint. Switching the website to https was a decision made by the directorate aimed at the positive image of the institution when maintaining a secure website. And BnF projects are about bringing traffic from search engines (see above).
The unchanged data at the SPARQL endpoint was not so much in focus because the endpoint is not much used and also they are looking at new API on top of that endpoint. Perhaps the http/https issue can be addressed in that layer.

Since there are several active users of the DNB Linked Data Service in the room they are asked: What happens to your system if DNB changed to HTTPS tomorrow? One wouldn't mind at all. Two would need advanced warning, but general attitude "changes are ok as technology advances. Just need to be announced". Users of the Zeitschriftendatenbank Linked Data Service are assumed to have a similar attitude but would need to be asked.

### Conclusion (emotion/tendency) of the group

Leans towards switching to https URIs as early as possible with advanced warning to data users.

# Materials

- Harry Halpin: "Semantic Insecurity: Security and the Semantic Web", October 2017 (gist: *switch now to https it's not too late yet and the web will be more secure immediately*)
- 2017 "The use of https: IRIs on the semantic web" thread on semantic-web@w3.org (*about https in vocabulary URI schemes in particular*, gist*: strong tendency to go for https*)
- Discussion page of the Wikidata community 2017-2018 (gist: *no change to https URIs. Possibly more due to the fact that the discussion subsided rather than an active decision*)
- W3C Blog post "HTTPS and the Semantic Web/Linked Data", 20 May 2016 (gist: *keep writing “http:” and trust in infrastructure and software to advance*)
- Discussion 2015 in the W3C Web Annotation Working Group (gist: *some statements for "don't use https URIs as identifiers" but discussion closed down for other, formal reasons*)
- W3C Design Issues, Tim Berners-Lee "Web Security - TLS Everywhere, not https: URIs", 2015 (gist: *the "s" breaks the web, use HTTPS but don't change the URI prefix*)

# More observations/inquiries results

- Wikidata has a 301 redirect from http to https URIs in place but retains http URIs as "Concept URIs", i.e. in RDF data.
- schema.org vocabulary retains http in the URIs, has a 301 redirect from http to https URIs and a triple in the RDF data connecting the two with schema:sameAs, e.g. `schema:CreativeWork schema:sameAs <https://schema.org/CreativeWork>` . (https://schema.org/CreativeWork.ttl, Status 2019-02-01) In the FAQs there is a statement that both, http and https are fine right now and that over time migration to https is intended (Status 2109-04-16)
- Library of Congress has so far decided not to switch URIs to HTTPS due to disruption but also lack of capacity to make a final decision. There are plans to review the decision and investigate further (Status February 2019)
- DOI URI prefixes are https://doi.org/, see http://www.doi.org/factsheets/DOIProxy.html#encoding
- ISNI URI prefixes are http://isni.org/isni/, see http://www.isni.org/how-isni-works
- ORCID URIs prefixes are https://orcid.org/, see https://support.orcid.org/hc/en-us/articles/360006897674-Structure-of-the-ORCID-Identifier
- [Update 2019-03-12:] German National Library (DNB) decided to switch all URIs in domain d- nb.info to https in October 2019
- The National Library of Sweden added new canonical "https" URIs for formal resources in the domains id.kb.se and libris.kb.se at the occasion of the introduction of their new infrastructure Libris XL. (The old "http" URIs are maintained using owl:sameAs in this new system. These will be issuing HTTP redirects to the new URIs as soon as possible.) (Status March 2019)

tbc