RDF Dataset Canonicalization

Scalable security for Linked Data

RDF

- RDF stands for Resource Description Framework
- Anything could be a resource. A resource might be…
  - real or fictional
  - concrete or abstract
  - digital or physical

RDF

- Resources are identified with IRIs (Internationalized Resource Identifiers)
- URLs are IRIs
- The IRI of the resource which represents FOSDEM on Wikidata:

  `http://www.wikidata.org/entity/Q475430`

RDF

- RDF statements describe resources and their relationships

```
<http://www.wikidata.org/entity/Q116786108>
<http://www.wikidata.org/prop/direct/P31>
<http://www.wikidata.org/entity/Q475430> .
```

- This statement is “FOSDEM 2024 is an instance of FOSDEM”

Collaborative knowledge graphs

- Wikidata: ~1500 million RDF statements
  - Covers anything ‘notable’ and ‘clearly identifiable’
  - Interoperates with OpenStreetMap to enrich information about sites of interest

https://www.wikidata.org/

Collaborative knowledge graphs

- MusicBrainz: ~100 million statements
  - Describes any published music and related data
  - Mostly RDF-compatible data model
  - Full RDF support (LinkedData) not yet realised

https://musicbrainz.org/

Institutional RDF datasets

- data.europa.eu is the official portal for data from EU member states
  - Thousands of individual datasets provided as RDF
  - RDF metadata about all ~1.5 million datasets

https://data.europa.eu/

Serialization of RDF

- Serialization formats make it easier for humans to read RDF data
  - XML/RDF
  - Turtle
  - JSON-LD
- Different formats don’t change the data at all

Serialization – a problem?

- Different serialization formats
- Dimensions of variation within formats:
  - Tabs or spaces, CR or CRLF, UTF-8 or UTF-16
  - Order in which statements are listed
  - Which IRIs are abbreviated

Serialization – a problem?

- Cryptographic signatures and integrity hashes work on a specific series of bits
  - Dataset in JSON-LD with spaces: d313cdbe...
  - Dataset in JSON-LD with tabs: cc52f7a2...
  - Dataset in Turtle: 8b80cc58...
- There are infinitely many ways to serialize the same RDF dataset, each with a different hash!

Canonicalization

The W3C RDF Dataset Canonicalization specification defines a **canonical serialization format** for RDF statements and an **algorithm to produce it** from data in other serialization formats

https://www.w3.org/TR/rdf-canon/

Open Standards

- W3C (World Wide Web Consortium) only develops specifications that everyone is free to use without restriction or royalties
- The RDF Dataset Canonicalization and Hash Working Group is one of 43 active working groups in the W3C

Verifiable Credentials

- Developed by the W3C Verifiable Credentials Working Group
- A standard for decentralized, privacy-respecting, offline credentials
- Likely to be the basis for EU personal ID cards
- Supports RDF Dataset Canonicalization

Implementations of rdf-canon

- Sophia (Rust, CeCILL-B Free Software License Agreement)
- rdfjs-c14n (TypeScript, W3C Software and Document license - 2023 version)
- rdf-canonize (JavaScript, BSD-3-Clause licence)
- rdf-canon (Rust, MIT licence)
- RDF::Normalize (Ruby, ‘Unlicense’)

Implementations of rdf-canon

JavaScript example with the rdf-canonize library....

```
canonize.canonize(nquads, {
  algorithm: 'RDFC-1.0',
  inputFormat: 'application/n-quads'
});
```

... where ‘nquads’ is your RDF dataset

Poison graphs

- Some RDF datasets have graph structures that are extremely slow to analyse
- Deliberate submission of such datasets can be a Denial of Service attack against tooling
- Mitigation is to place limits on computation when performing RDF Dataset Canonicalization

Poison graphs

- Low risk in non-malicious contexts:

…in our evaluation we demonstrate that there indeed exist difficult synthetic cases, but we also provide results over 9.9 million RDF graphs that suggest such cases occur infrequently in the real world, and that both canonical forms can be efficiently computed in all but a handful of such cases.

(2017) Canonical Forms for Isomorphic and Equivalent RDF Graphs: Algorithms for Leaning and Labelling Blank Nodes, by Aiden Hogen

Contribute

- Implementations are always appreciated!
- Future improvements will be needed to support RDF 1.2
- Ask to become an W3C Invited Expert (for independent individuals) or register as a W3C Member (for organisations)

Thank you for listening!

- Sebastian Crane
- seabass-labrax@gmx.com
- @seabass:fosdem.org