Resuscitating the Semantic Web

Spencer Killen
https://sjkillen.ca

About Me

- Doing knowledge representation research at UoA
  - Computer languages for human knowledge
- The Semantic Web overlaps with my research

A Technical Writing Problem

- Thesis of papers

A Technical Writing Problem

- Thesis of papers
- Typical math paper structure:
  - Preliminaries
    - Redefine definitions from other papers
    - Adjust definitions, use different notation, express in a equivalent, different manner
  - The technical contributions

A Technical Writing Problem

- Thesis of papers
- Typical math paper structure:
  - Preliminaries
    - Redefine definitions from other papers
    - Adjust definitions, use different notation, express in a equivalent, different manner
  - The technical contributions

Consolidating papers is **really** hard

Consolidating Papers

- First, find overlap in the preliminaries sections

Consolidating Papers

- First, find overlap in the preliminaries sections
- Rectify differences
  - E.g. paper one calls thingamajig a thingamabob, but paper two calls them doohickeys.
  - The combined work should use the same terms throughout for the same object
    - It should be consistent

Consolidating Papers

- First, find overlap in the preliminaries sections
- Rectify differences
  - E.g. paper one calls thingamajig a thingamabob, but paper two calls them doohickeys.
  - The combined work should use the same terms throughout for the same object
    - It should be consistent
- A system to make this easier?
  1. Verify that vocabulary is consistent (only one word per thing)
  2. Make sure that every technical concept is defined somewhere
  3. Enhance document with links to definintion

What is the Semantic Web?

- Internet of unstructured documents (current)
- Internet of knowledge (semantic web)

A Brief History of Internet "Versions"

- Web 1.0
  - Website, basic browsing, hyperlinks

A Brief History of Internet "Versions"

- Web 1.0
  - Website, basic browsing, hyperlinks
- Web 2.0
  - Dynamically loaded content
  - Infinite scroll, more interactivity, etc.

“Versions” are Largely Marketing Slop

- Hundreds of “Web 2.0” books sold

“Versions” are Largely Marketing Slop

- Hundreds of “Web 2.0” books sold
- But... versions do signal a push for
  - an epoch

The Semantic Web: Web 3.0

- 2010s, people pushed the Semantic Web as Web 3.0
  - Mainly academics
  - Low industry adoption

The Semantic Web

- A phrase coined by Tim Berners-Lee
  - Who is credited for “inventing the internet”

The essential property of the World Wide Web is its universality. The power of a hypertext link is that ”anything must be able to link to anything”

– Tim Berners-Lee

The Semantic Web

- In Web 1.0/2.0 documents of structured information link to each other.
  - But the documents don't *mean* anything except to a human reader.

```
<div>
    <span>1-780-468-2639</span>
    <span><a href="https://sjkillen">
        Spencer Killen
    </a></span>
    <span>sjkillen@ualberta.ca</span>
</div>
```

Level 0: Data Annotation

- Documents describe what they are

```
<div typeof="Person">
    <span property="CellPhone">
        1-780-468-2639
    </span>
    <span property="Website">
        <a href="https://sjkillen">
            Spencer Killen
        </a>
    </span>
    <span property="Email">
        sjkillen@ualberta.ca
    </span>
</div>
```

Level 0: Data Annotation

- Meaning ascribed to different sections of the document
- There's an automated way of interrogating the web resource:
  - Who is the website about?
    - One person, Spencer
  - What is Spencer's email?
    - sjkillen@uaberta.ca

The Good and the Bad

- Website is **more** machine readable
- Easier for bad corpos to mine information
  - But companies will use LLMs to mine this info with or without annotation
- Good actors:
  - Application interoperability
  - E.g. Very easy to build an application that looks through pages you've browsed
    - "Makes a list of the people I've read about online" or
    - Poll https://sjkillen.ca periodically and update Spencer's contact info

Level 1: Distributed Authority and Ontologies

- An information either asserts authority:
  "the sky is blue, trust me bro"
- or cites a source: $e = mc^2$ (Einstein 1905)

Level 1: Distributed Authority and Ontologies

- Ideally, anything and everything should be cited
  - Current methods of citation are highly limited
    - We reference a book, article, etc and expect the reader to find the relevant parts.
    - More granular systems exist (e.g. citing specific diagrams, bible verses, etc)
    - But these systems cite some structural part of the source material, not a semantic one.

A Contrived Example

Example

Writing a fanfic which combines Linux lore with Lord of the Rings.
Richard Stallman is a Hobbit character.
A reference to the Stallman Hobbit needs to cite both Linux lore and LOTR.
How will people know which idea comes from which source?

```
<span typeof="en.wikipedia.org/Richard_Stallman">
    Richard
</span>
was a 3 foot tall
<span typeof="lotr.com/Hobbit">
    Hobbit
</span>
```

- Normal citations happen once, but we can cite every usage

Authority Delegation

- Now Wikipedia and lotr.com are the authoritative definitions of Stallman and Hobbits, respectively

Modeling Concepts, Relations, and Webs

- Our citation doesn't describe the relation our character has to the real person our LOTR books.

```
<div target="OurProtagonist">
    <span property="schema.org/isBasedOn">
        <span href="en.wikipedia.org/Richard_Stallman">
                Richard
        </span>
    </span>
    was a 3 foot tall
    <span typeof="lotr.com/Hobbit">
        Hobbit
    </span>
</div>
```

Schema.org

We Can Define Our Own Concepts Too!

- Someone can write a fanfic of our fanfic to cite and incorporate a new creature we define
- Someone browsing for stories can search to find all works that incorporate our creature
  - If usages are self-reported, therefore not susceptible to AI slop errors.

Level 3: Semantic Services

- A website can use these technologies to report the services it provides
  - E.g. a pi digit calculator
  - An imperial-to-metric converting service

The Premature Death of the Semantic Web

"A technology dies first when people stop using it, and then again when its name is lost. "

– Made up

The Premature Death of the Semantic Web

> *"A technology dies first when people stop using it, and then again when its name is lost. "*
>
> *– Made up*

- "Crypto/Blockchain is now Web 3.0"!

The Premature Death of the Semantic Web

> *"A technology dies first when people stop using it, and then again when its name is lost. "*
>
> *– Made up*

- "Crypto/Blockchain is now Web 3.0" !
  - The name has been stolen from the Semantic Web

The Arduous Knowledge Engineering Barrier

- Implementing the semantic web takes two types of experts working in tandem:
  - Experts in web technologies (knowledge engineers)
  - Experts in the domain (e.g., Linux lore expert or LOTR expert)
- This is why semantic web technologies never saw widespread adoption

Reclaim the Web from the LLMs!

- Large language models are stealing the web from us
  - Automatically extracted data from websites is served in a chat client
  - Websites aren't visited by people, just LLM scrapers

Reclaim the Web from the LLMs!

- Large language models are stealing the web from us
  - Automatically extracted data from websites is served in a chat client
  - Websites aren't visited by people, just LLM scrapers
- Fighting back?
  - Use LLMs to allow the domain expert to "be the knowledge engineer"
    - Ensures automatic data extraction is audited by an expert, unlike current LLMs
    - Data is accessible and reliable
    - Greatly decreases the cost of semantic web tech adoption
  - Build more tools that rely on the Semantic Web

Search Over Synthesis

- LLMs synthesize information to answer queries
  - Slop is often wrong
  - Synthesized sources often wrong
- The semantic web can make search as powerful as chatbots
  - And be correct

Semantic Web Technologies

- RDF
  - RDFa
    - Used in examples. Annotate HTML with relations
- OWL
  - Extends RDF with description logic
- RIF
  - Extends RDF with intuitionistic rule logic

The Semantic Web in the Wild

- ActivityPub

How to Get Started

1. Learn basic HTML
   - Only very basic HTML is needed.
2. Learn RDFa Core
3. Annotate a document using schema.org