Web Annotation

Anıl Selim Sürmeli
Bogazici University
Software Engineering Department
Istanbul, Turkey
anil.surmeli@gmail.com

***Abstract*— we have limited interaction with the content on a web site: reading the text, viewing an image, clicking on links and creating bookmarks. Although we can participate and create communities in an application after Web 2.0, people still bordered to explain their thoughts in a comment box generally below the content, no more. That's why web annotation has become trendy topic in software development world as people recognized that the current one-directional information sharing system on Web is no longer sufficient. This research is going to summarize what web annotation and web annotation data model is, history of web annotation standards, modern tools around the web and future of the web annotation.**

I. Introduction

Formally, annotation can be defined as a value adding note or marking that is linked to an extant information object, representing a record of interaction between the reader and the information object (McMullen, 2005). That means, user can feel like taking a note in a text book if web annotation is enabled for the web content, reader can highlight the text, comment only one phrase in a paragraph, mark a part of picture and so on.

In fact, some digital tools have capability of annotation in the content. Microsoft Word allows users to track changes when they edit a document [1], user can also comment, underlie or highlight the content in Adobe Reader. However, in web world, there are limited options for annotating the content. Some plugins enhances web browsers with annotation functionality, such as Genius, Hypothes.is, A.nnotate, Awesome Highlighter, Blerp, Bounce and Quickfox. Also there are blogging platforms that has annotation capabilities by default like Medium.com.

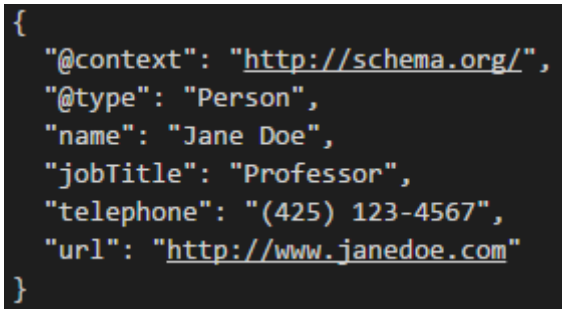
II. Brief History of Web Annotation

Web annotation was implemented by various tools and individual web applications by using their own logic and data structure. After 2012, Open Annotation Working Group has drafted Open Annotation Core Data Model and specified an approach for creating associations between related resources [2]. Many tools such as Annotator.js and Hypothes.is implemented the standard [4]. Open annotation community have decided to end the project in 2013 in order to help the work of the W3C Web Annotation Working Group [5]. In 2016, W3C's Web Annotation Working Group has drafted the Web Annotation Data Model. In the newer specification, an annotation is defined as a web document and identified with a URI of its own, moreover, an annotation can have an annotation too.

III. WEB ANNOTION DATA MODEL

JSON-LD is human readable, JSON based and modern linked data structure that is gaining popularity recent years. Technology giants such as Google and Microsoft supports JSON-LD format in their products, e.g. Gmail. What is interesting in JSON-LD is that it has "@context" field, which states some kind of namespace for the vocabulary. JSON-LD fields are all defined in the vocabulary given in "@context" field.

JSON-LD in extended format.



JSON-LD in compacted format.

Compacted and expanded representations are actually same. Using JSON-LD Processor, developers easily convert the representations.

Official JSON-LD processor supports expanded, compacted, flattened, framed, N-Quads, Normalized, signed with RSA and Signed with Bitcoin versions according to the official web site.

Using linked data with JSON-LD allows search engines to understand your data.

As an example, when googling a famous musician such as Ritchie Blackmore, it can be seen that there is a summary card near the search result, summary of the result including Ritchie Blackmore's hometown, band information, age, family information and etc.

So why the summary card does not appear when searching a random person? That is probably because the person is not represented in linked data on web.

JSON-LD can also be used in converting semantically equal objects. Think about you are going to implement a social media platform and you will enable Facebook and Twitter authentication for your application. Your user object contains “name” field whereas Facebook API describes it as “userName”, also Twitter API represents the same data as “accountName”.

In traditional approach, you should convert third party API data to your domain object by equaling Facebook’s “userName” with your “name”. Same approach happens for Twitter API.

If you use numerous third party APIs, implementation is going to be complicated by converting tens of fields.

However, using JSON-LD structure, contexts of third party APIs can be easily convert to your context using JSON-LD processor, that is, APIs use a user object described in schema.org, they can easily understand each other without unnecessary convert process.

Another and the main advantage is that, linked data can easily have associated between some other linked data on web, by just giving a link for a vocabulary defined property.

Schema.org

Schema.org is an initiative to "create and support a common set of schemas for structured data markup on web pages" launched by Google, Bing and Yahoo [6]. Schema.org can be used in Microdata, RDFa and JSON-LD formats as vocabulary source and recognized by
search engine spiders.

W3C Web Annotation Data Standard

For annotations, we use linked data and JSON-LD format to represent the annotation data as annotations are resources on web. You can use an annotation for multiple targets, annotate some other annotation and so on.

Web Annotation Working Group is part of the W3C Digital Publishing Activity. The group has created three specifications which are:

- Web Annotation Data Model
- Web Annotation Vocabulary
- Web Annotation Protocol

Simple annotation data representation based on JSON-LD:

```
{
  "@context": "http://www.w3.org/ns/anno.jsonld",
  "id": "http://example.org/anno1",
  "type": "Annotation",
  "body": "http://example.org/post1",
  "target": "http://example.com/page1"
}
```

According to the standard, an annotation must have all the fields above except body, that can be null representing the highlight on target.

IV. The Future of the Web Annotation

Web annotation has become interesting topic in recent years. Many academicians, students, dynamic organizations and innovative corporations use Web Annotation [6]

The future is bright as the standard is shaped and sharpened by W3C using latest linked data format, JSON-LD.

Semantic Annotation

Machines can easily refer semantic annotations that are tagged by user. With tagging, annotation targets become a resource for an information which enriches content by linking background information to extract concepts.

Example case on text identification [7]:

“Rome was the center of the Roman Empire and there were over 400.000 km of roman roads connecting the provinces to Rome.”

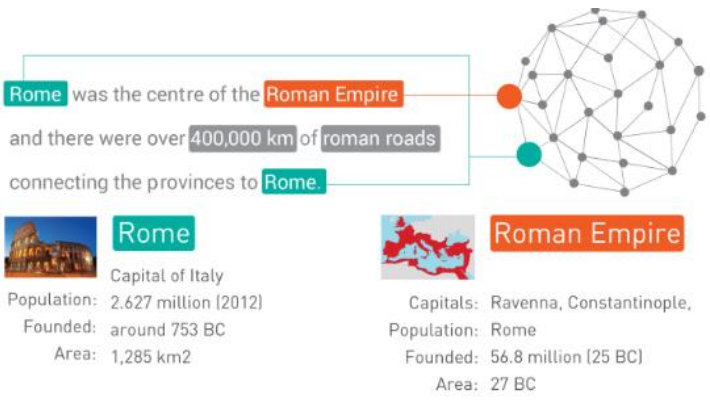
Text Analysis:

“Rome was the center of the Roman Empire and there were over 400.000 km of roman roads connecting the provinces to Rome.”

Important concept such as proper nouns can be extracted from target annotation. Algorithms split sentences and keywords.

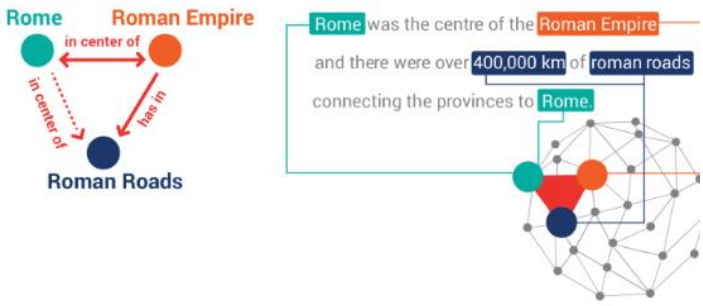
Concept Extraction:

Machine learning algorithms work for entity classification, the most important stage of semantic web annotation [7]:



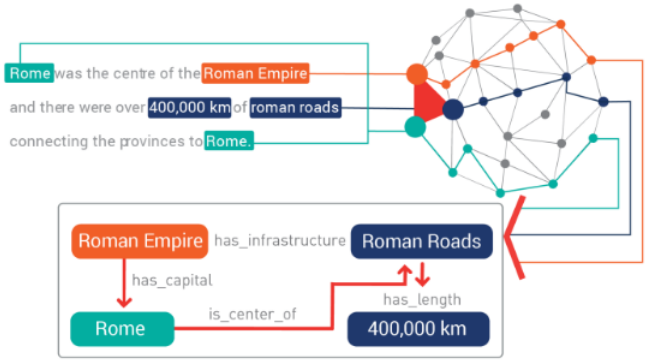
Relationship Extraction:

Relationships between entities are extracted using structured annotation data and linked data engines.



Indexing and storing in a semantic graph database

All the annotations are indexed and persisted in annotation database, smart queries and optimized reflections will be run using graph algorithms.



V. CONCLUSION

We use annotations for decades by underlying, highlighting and taking notes on a hard copied resource. There are also many technologies for us to annotate a target on a digital resource such as pdfs, and word documents.

Today, web annotation popularity increases by the help of useful plugins such as A.nottate, Genius and Hypothes.is.

Before the web annotation standard, many applications stored and represented annotation data using their own algorithms and representation models. Open Annotation Group was the first community publishing a standard. Web Annotation Working Group is now the author of current model.

JSON-LD is new linked data format, which also improved version of metadata and RDFa.

The future of web annotation is bright and today, most of the web giants support web annotation. Official documentation of Microsoft frameworks allows web annotation on the resource, also such blogging platforms like Medium and Ghost have annotation features.

REFERENCES

[1] http://onlinelibrary.wiley.com/doi/10.1002/meet.14504201151/full
[2] http://www.openannotation.org/spec/core/20120328.html
[3] https://www.w3.org/community/openannotation
[4] https://hypothes.is/blog/supporting-open-annotation/
[5] http://www.openannotation.org
[6] https://schema.org/docs/faq.html
[7] http://ontotext.com/knowledgehub/fundamentals/semantic-annotation/