Human Factors in Software and Systems Engineering, Vol. 94, 2023, 95–103
https://doi.org/10.54941/ahfe1003774

Evaluating Embedded Semantics for Accessibility Description of Web Crawl Data

Rosa Navarrete, Diana Martinez-Mosquera, Lorena Recalde, and Marco Aguirre

Departamento de Informática y Ciencias de la Computación, Escuela Politécnica Nacional, Quito, Ecuador

ABSTRACT

The growing of the Web joined with the pandemic's conditions, have motivated an increasing number of people to use it for communication, learning, work, commerce, business, entertainment, and more. So, the need for equitable web access for people with disabilities is mandatory. The web searching is a striving activity, particularly for these vulnerable group of people because they need to examine many results to find those that meet their accessibility requirements. The embedded semantics is a novel approach to improve results provided by search engines getting a better user experience in web searching. In this research, we processed web crawl information, released in 2021 year, which contains millions of websites, to analyze the use of embedded semantics, particularly, annotations with accessibility properties. The analysis implied the configuration of a big data processing platform and produced quantitative results concerning accessibility descriptors in web domains and other kinds of descriptors that appear jointly with accessibility properties. Furthermore, we obtained a normalized dataset stored in MongoDB for further analysis.

Keywords: User experience, Data mining, Structured semantic markup, Educational resources, JSON-LD

INTRODUCTION

People globally use the Web as primary platform for communication, learning, work, commerce, business, entertainment, and more. Therefore, accessing the web must be equitable for all people, regardless of their disabilities, which implies addressing web accessibility issues. The World Wide Web Consortium (W3C), the leading organization responsible for ensuring the growth of the social value of the Web, establishes standards, protocols, and recommendations to improve the reach extent of web content for people. For instance, Web Content Accessibility Guidelines (WCAG) promote the achievement of web accessibility. Furthermore, other W3C recommendations foster embedded semantics into the web content to help browsers build a machine-readable data structure aiming to produce an enriched description in search results supporting people to find the right content for their queries and, consequently, improving user experience (Guha et al., 2015). Searching for specific web content is especially striving for people with disabilities because they might require examining many search results before finding some content that matches their accessibility requirements (Aizpurua et al., 2016). If embedded semantics communicate the accessibility properties of content, search will be more productive for everyone, but even more so for people with special needs.

Embedded semantics requires two components, a vocabulary, and an encoding format. The Schema.org vocabulary has experienced a great growth and encompasses plenty of descriptors for each type of web information (Yu, 2014), including the set of descriptors for accessibility characteristics information (Schema.org, 2022). Regarding the encoding format, JSON-LD is the latest W3C recommendation due to its ability to make JSON data interoperate at Web-scale. It provides a quickly transforming for Linked Data format and is simple enough to be read and written by people (W3C, 2021).

This research conducted a quantitative analysis of the semantics embedded in web content by processing a dataset obtained from millions of web crawl data released in the year 2021. The dataset included web content from distinct provenance and purposes at a global scale. In this web content, embedded annotations are done using JSON-LD script that uses Schema vocabulary elements to communicate inherent semantics.

The analysis defined how the accessibility descriptors are used jointly with other classes and properties to describe the web information on personal blogs, organizations, events, educational content, universities, commerce, sports, medicine, entertainment, among others. The results provided a perspective of the awareness of accessibility issues in the different fields of the Web. The processing was performed on collected zip files that contain over three hundred million records. This analysis was performed using techniques applied to big data, such as key-value modelling with Python for processing and a NoSQL database such as MongoDB for storage.

The contributions of this research are twofold. In the first place, the analysis of the interest in the Web in using accessibility descriptors in embedded semantics. The quantitative results enable us to appreciate the concern for equity and inclusion that is made visible through accessibility issues in different entities, according to the web domains. In addition, these results reveal how the W3C recommendation of embedded semantics is being adopted to create a more organized and better-documented Web. Secondly, processing the raw dataset results in a new normalized dataset in JSON format with information about domains, web content types, and properties associated with the accessibility descriptor. This new dataset will be available for further analysis of embedded semantics.

The remainder of the paper is organized into the following sections: Related Works, which presents a summary of relevant works; Materials and Method, which explains the provenance of the dataset for the analysis and the method used to obtain the results; Results, which details and discusses the findings; and finally, Conclusions and future work expose final ideas.

RELATED WORKS

Some research works have addressed embedded semantics and related issues. The use of embedded semantics to enhance results on a web search was presented by (Haas et al., 2011). In this work, the authors propose the use of mechanisms to translate web page metadata into search result displays in terms of terms of users' search experience. The findings show that users appreciate refined search results with content such as images and other interactive elements. In addition, how Google obtains enriched search by exploring embedded semantics is presented in (Ohshima and Toyama, 2018). On the other hand, the embedded semantics to support device interactions is exposed in (Mayer and Basler, 2013), and the embedded semantics applied to numerical series data is developed in (Hossayni et al., 2018). The application of Schema's vocabulary for markup in the Cultural Heritage domain is exposed in (Freire et al., 2018). In the same way, the use of Schema's vocabulary in metadata is presented in (Gregory et al., 2020) with application to the discovery of data needed for research. The importance of embedded markup to accuracy in describing search results is presented in (Bakhouyi et al., 2020), while the discoverability issues in web searching using embedded semantics are presented in (Kraker et al., 2021). A guide for best practices in embedded structured data have been published by (Wu et al., 2021).

Several research works have been carried out concerning the use of embedded semantics in specific fields; as instance, in the educational field we have (Navarrete et al., 2019); (Bakhouyi et al., 2020); (Recalde et al., 2021); (Recalde et al., 2022).

However, despite the importance of the topic, in the literature review, we do not find, so far, any research focused on the use of accessibility descriptors for embedding semantic annotations. For this reason, it is valuable to know, through the findings in the present work, how accessibility is considered for web admins and developers in the world concerning the use of this alternative to produce structured data.

MATERIALS AND METHOD

The currently accepted vocabulary of Schema.org that represents accessibility properties is depicted in Table 1. The names of properties are written in cursive format, as usual in the Schema context. The accessibility properties complement and document different kinds of descriptors for Things, Creative Works, Organization, Person, Place, Product, Event, and more.

The dataset for analysis is provided by Web Data Commons (WDC), an organization that releases extracted data from Common Crawl (CC), the largest web corpus available to the public. The dataset was released in October 2021, containing structured data with semantic annotations made with JSON-LD (WDC, 2021). The raw data are presented as a compressed set with 6240 gzip files, with around 100MB each.

Every.gz file has many records, and each record represents an N-quad that means a triple RDF in a single line; it can configure a valid RDF (subject, predicate, object), a label or blank node, or an IRI (URI or URL) to identify the web page where the triple was extracted. The processed data is presented in Table 2.

Table 1. Set of accessibility properties of schema.org.

Property	Purpose
accessibilityAPI	Compatibility with the referenced accessibility API.
accessibilityControl	Input methods that are sufficient to control the described resource fully.
accessibilityFeature	Content features of the resource, such as accessible media, alternatives, and supported enhancements for accessibility.
accessibilityHazard	A characteristic of the described resource that is physiologically dangerous to some users.
accessMode	The human sensory perceptual system or cognitive faculty through which a person may process or perceive information.
accessModeSufficient	A list of single or combined *accessModes* that are sufficient to understand all the intellectual content of a resource.
accessibilitySummary	A human-readable summary of specific accessibility features or deficiencies, consistent with the other accessibility metadata.

Table 2. Values of corpus and datasets.

Subject	CC Corpus	WDC dataset
Number of HTML pages	2,500,000,000	900,000,000 (36% of the HTML pages in CC)
Number of domains	32,884,530	9,650,571 (29.3% out of the total domains in CC)
Size of data	54 TB (compressed)	653 GB (compressed)
.gz files		6240
N-quads processed		38,420,581,511

A total of 6240 zip files with embedded JSON-LD data were downloaded and contain over 38 billion N-quads. The architecture deployed for analyzing accessibility data is presented in Figure 1.

These data were processed into a data lake with 32GB of RAM, 2.5TB of storage, and 10 CPUs. In addition, they were stored in a collection of JSON documents in MongoDB. The data cleanup process included the following steps:

1. Decompress the gzip files.
2. Find the schema.org tag in the rows.
3. Verify if that is a valid N-quad.
4. Split the N-quad.
5. Verify whether every N-quad contains accessibility properties.
6. Store valid quads with accessibility properties in a MongoDB collection with the following records: N-quad identifier, N-quad number, property, value property, and domain.
7. Store non-valid N-quads in a MongoDB collection named logs.

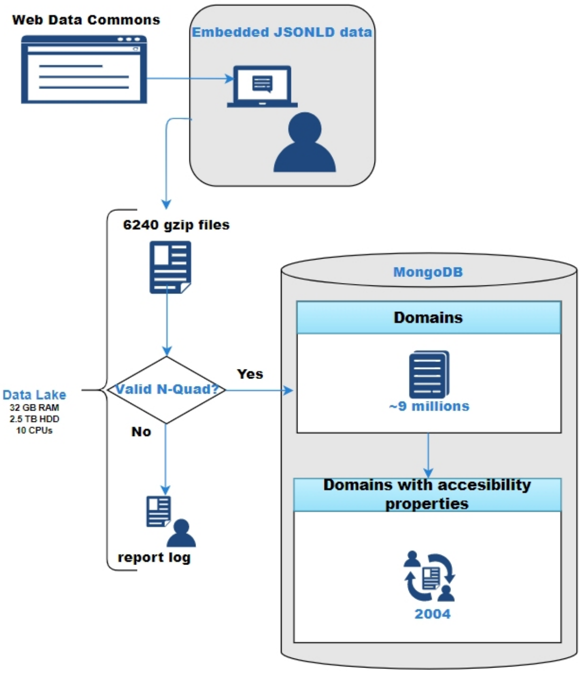

Figure 1: Architecture for JSONLD data processing.

A new dataset containing normalized data was generated with information about domains, types of web content, and properties associated with the accessibility descriptor. The collection and storage layers were implemented in MongoDB and are available for new consumption in the cloud.

RESULTS AND DISCUSSION

As a result of this research, we found that, out of 9,650,571 domains obtained within the dataset, only 2004 contained accessibility properties, barely 0.02%. Figure 2 indicates the number of domains that include accessibility properties. The top three most representative properties were *accessibilityFeature*, *accessMode*, and *accessModeSufficient*; these properties provide significant information about accessibility, so they are the most used. The remaining properties did not have a considerable presence in domains.

Figure 3 shows the outcome of other properties used in domains that also used accessibility properties. Only *Organization* is a "superior class" property found jointly with accessibility properties. A "superior class" is a schema vocabulary term representing something that can be described with more attributes or descriptors.

Considering the N-quads processed, Figure 4 shows the number of N-quads for each accessibility property. We found that *accessMode* is the property most used, representing almost 65% of total N-quads that present accessibility properties. This finding corroborates the importance of this

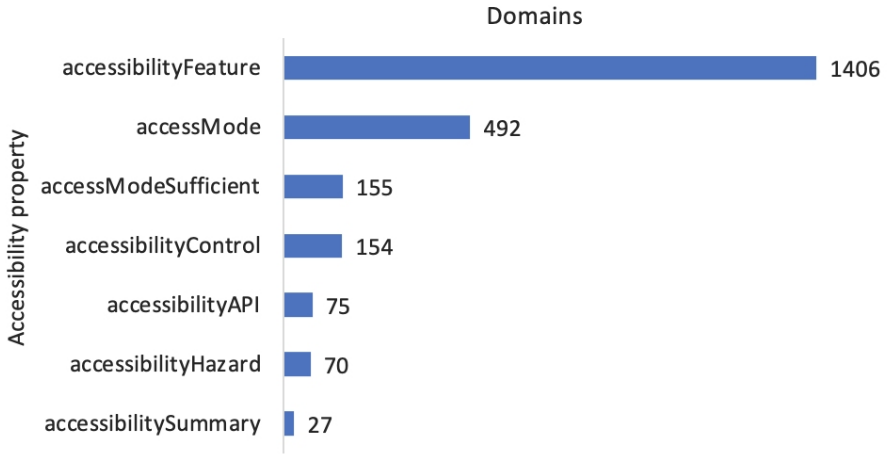

Figure 2: Number of domains with accessibility properties.

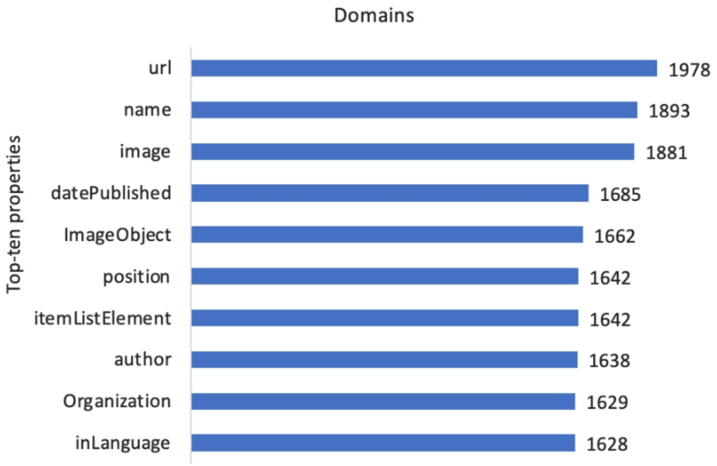

Figure 3: Number of domains with co-occurrence of accessibility properties and others.

property which declares the type of perception required about the information (auditory, tactile, textual, or visual). Also, *accessibilityFeature* appears as the second property; because it conveys information about specific accessibility characteristics, such as caption or transcript for videos or sign language availability as an alternative to audio.

The number of N-quads of “superior class” properties found in domains that use accessibility properties is presented in Figure 5. The class *WebSite*, *CreativeWork*, and *WebPage* are prevalent. Nevertheless, *Website* and *WebPage*, as well as other “superior class” properties such as *article*, *course*, and *book*, are subclasses of *CreativeWork*. In some cases, the “superior class” *CreativeWork* has probably not been described with specific properties. These results confirm the purpose of embedded semantics to provide information about the entire element more than a particular one.

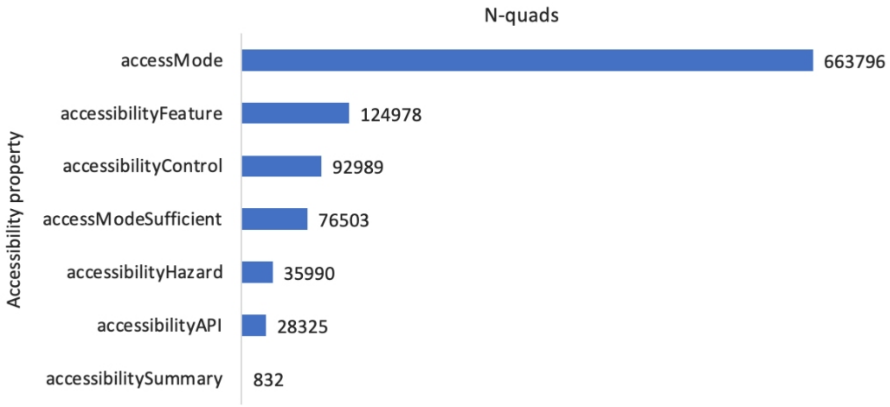

Figure 4: Number of N-quads for each accessibility property.

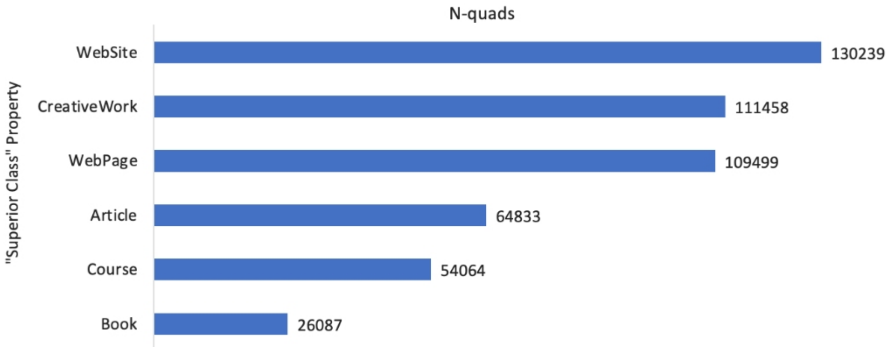

Figure 5: Number of N-quads for "superior class" in domains with accessibility properties.

Table 3. Types of domains with accessibility properties.

Description	Root zone	Number
VeriSign Global Registry Services	.com	1217
Public Interest Registry (PIR)	.org	87
Stichting Internet Domeinregistratie Nederland (SIDN)	.nl	55
DENIC eG	.de	54
Viet Nam Internet Network Information Center (VNNIC)	.vn	49
VeriSign Global Registry Services	.net	48
Association Française pour le Nommage Internet en Coopération (AFNIC)	.fr	37
IIT - CNR	.it	34
Nominet UK	.uk	33
Ministry of Information and Communications Technologies (MinTIC)	.co	18

Finally, concerning the typology of domains that use accessibility properties, we grouped them attending the root zone defined by the Internet Assigned Numbers Authority (IANA). Table 3 exposes the number of the domains for each root zone found in the analysis. As can be seen, .com domains are by far the most representative; this fact suggest that commerce websites are interested in use the embedded semantic to deliver search information aimed to people with special needs.

CONCLUSION AND FUTURE WORK

Considering the great diversity of people using the Web, including people with special needs, and regardless of the purpose of web domains, it is essential to take advantage of all tools that enhance user experience. In such context, embedded semantics enable search engines and user agents to parse and convey better and complete search results descriptions (Research Data Alliance, 2020). Incorporating accessibility descriptors enriches the information transmitted to people with special needs (Navarrete et al., 2019).

Contrary to what has been exposed, the results of the data analysis show a minimal use of accessibility descriptors in the websites included in the data. Even though web pages have embedded vocabulary, developers and web admins do not seem to know about this set of accessibility descriptors.

As relevant results of our research, we found that the most used accessibility properties are *accessMode*, and *accessibilityFeature*, because these properties communicate substantial information regarding accessibility characteristics of the web content. Also, we found that "superior classes" more frequently used in domains that also incorporated accessibility properties are *WebSite* and *WebPage*, confirming the decision to embed semantics that describe preferably the entire web content more than the specific elements.

Regarding the domains, we found that, .com domains are prevalent for including accessibility descriptors in embedded semantics. That indicates that commercial organizations are most concerned about accessibility issues.

In future work, we propose to make a multi-year analysis considering the last three years in crawling data to detect if, gradually, the use of accessibility descriptors is incremented. Also, it is essential to discover if web developers and publishers do not know the existence of these descriptors or if they consider that the use is not convenient, probably because they are not meaningful to the delivery of information.

REFERENCES

Aizpurua, A., Harper, S. and Vigo, M. (2016). Exploring the relationship between web accessibility and user experience. *International Journal of Human-Computer Studies*, 91(), pp. 13–23.

Bakhouyi, A., Dehbi, R., Banane, M. and Talea, M. (2020). A Semantic Web Solution for Enhancing the Interoperability of E-Learning Systems by Using Next Generation of SCORM Specifications. *Advances in Intelligent Systems and Computing*, 1102 AISC, pp. 56–67.

Freire, N., Charles, V. and Isaac, A. (2018). Evaluation of Schema.org for Aggregation of Cultural Heritage Metadata. *The Semantic Web. ESWC 2018. Lecture Notes in Computer Science,* 10843() . Springer, Cham, pp. 225–239.

Gregory, K., Scharnhorst, A. and Wyatt, S. (2020). Lost or Found? Discovering Data Needed for Research. *Harvard Data Science Review*, 2(2).

Guha, R., Brickley, D. and MacBeth, S. (2015). Schema.org: Evolution of Structured Data on the Web: Big data makes common schemas even more necessary. *Queue*, 13(9), pp. 10–37.

Haas, K., Mika, P., Tarjan, P. and Blanco, R. (2011). Enhanced results for web search. *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 725–734.

Hossayni, H., Khan, I. and Kaed, C. (2018). Embedded Semantic Engine for Numerical Time Series Data. 2018 Global Internet of Things Summit, pp. 1–6.

Kraker, P., Schramm, M. & Kittel, C. (2021). Discoverability in (a) crisis. *ABI Technik*, 41(1), pp. 3–12.

Mayer, S. and Basler, G. (2013). Semantic metadata to support device interaction in smart environments. *UbiComp 2013 Adjunct - Adjunct Publication of the 2013 ACM Conference on Ubiquitous Computing*, pp. 1505–1514.

Navarrete, R., Montenegro, C., Recalde, L. and Lujan-Mora, S. (2019). Analyzing embedded semantic with JSON-LD and microdata for educational resources in large scale web datasets. 6th Annual Conference on Computational Science and Computational Intelligence, CSCI 2019, pp. 1133–1138.

Ohshima, T. and Toyama, M. (2018). SDC: Structured data collection by yourself. *ACM International Conference Proceeding Series, (March)*, pp. 16–18.

Recalde, L., Navarrete, R. and Correa, L. (2022). A Framework for data mining of structured semantic markup extracted from educational resources on University websites. AHFE 2022 Open Access, Usability and User Experience, pp. 482–489.

Recalde, L., Navarrete, R. and Pogo, F. (2021). Making Open Educational Resources Discoverable: A JSON-LD Generator for OER Semantic Annotation. *Eighth International Conference on eDemocracy & eGovernment (ICEDEG 2021)*, pp. 182–187.

Research Data Alliance (RDA) WG, 2020. Guidelines for publishing structured metadata on the Web V2.0, [Online] Available at: https://www.rd-alliance.org/ [Accessed November 2022].

Schema.org, 2021. Organization of Schemas. [Online] Available at: https://schema.org/docs/schemas.html [Accessed November 2022].

W3C, 2021. JSON-LD 1.1: A JSON-based Serialisation for Linked Data. [Online] Available at: https://w3c.github.io/json-ld-syntax/ [Accessed November 2022].

Wu, M., Juty, N., RDA Research Metadata Schemas WG, Collins, J., Duerr, R., Ridsdale, C., Shepherd, A., Verhey, C., and Castro, L. J. (2021). Guidelines for publishing structured metadata on the Web V3.1. *Research Data Alliance*.

Yu, L. (2014). A Developer's Guide to the Semantic Web, Berlin Heidelberg, Springer-Verlag.