future internet

MDPI

Article

Estimation on the Importance of Semantic Web Integration for Art and Culture Related Online Media Outlets

Andreas Giannakoulopoulos [1,*], Minas Pergantis [1,*], Nikos Konstantinou [1], Alexandros Kouretsis [1], Aristeidis Lamprogeorgos [1] and Iraklis Varlamis [2]

[1] Department of Audio and Visual Arts, Ionian University, 7 Tsirigoti Square, 49100 Corfu, Greece; nikoskon@ionio.gr (N.K.); akourets@gmail.com (A.K.); a18labr@ionio.gr (A.L.)
[2] Department of Informatics and Telematics, Harokopio University of Athens, 70 El. Venizelou Str., 17676 Athens, Greece; varlamis@hua.gr
* Correspondence: agiannak@ionio.gr (A.G.); a19perg6@ionio.gr (M.P.)

Abstract: Since the dawn of the new millennium and even earlier, a coordinated effort has been underway to expand the World Wide Web into a machine-readable web of data known as the Semantic Web. The field of art and culture has been one of the most eager to integrate with the Semantic Web, since metadata, data structures, linked-data, e.g., the Getty vocabularies project and the Europeana LOD initiative—and other building blocks of this web of data are considered essential in cataloging and disseminating art and culture-related content. However, art is a constantly evolving entity and as such it is the subject of a vast number of online media outlets and journalist blogs and websites. During the course of the present study the researchers collected information about how integrated the media outlets that diffuse art and culture-related content and news are to the Semantic Web. The study uses quantitative metrics to evaluate a website's adherence to Semantic Web standards and it proceeds to draw conclusions regarding how that integration affects their popularity in the modern competitive landscape of the Web.

Keywords: semantic web; media; art; culture; quantitative analysis; internet statistics; world wide web

Citation: Giannakoulopoulos, A.; Pergantis, M.; Konstantinou, N.; Kouretsis, A.; Lamprogeorgos, A.; Varlamis, I. Estimation on the Importance of Semantic Web Integration for Art and Culture Related Online Media Outlets. *Future Internet* **2022**, *14*, 36. https://doi.org/10.3390/fi14020036

Academic Editor: Rafael Valencia-Garcia

Received: 31 December 2021
Accepted: 17 January 2022
Published: 24 January 2022

1. Introduction

The Semantic Web, as a means to structure and disseminate data through machine-readability [1], is very important in the fields of art and culture and a cornerstone of digitized art and cultural heritage collections around the globe [2]. It is also a useful tool in the field of contemporary journalism, displaying enormous potential for information gathering, filtering and dissemination [3,4]. Since art and culture themselves are often the subject of journalistic content, the usage of Semantic Web technologies especially in media outlets that focus on these specific fields presents a very interesting landscape for research.

In the study presented in this article, the researchers proceeded to get a thorough glimpse at the landscape of Semantic Web information provided by art and culture-related websites with an added focus on the reportorial or journalistic aspects of this information. In order to do so, a variety of relevant websites were identified using a process that involved both automated and expert manual selection. These identified websites were then perused by an automated crawling algorithm and metrics about their integration of Semantic Web technologies were collected. The specific metrics selected were based on the researchers' expertise in the field of the Semantic Web and its application in art and cultural heritage and are presented in detail in the methodology section. Moreover, through this expertise an integration rating system was devised and is presented in detail. The information collected was further analyzed through means of not only traditional statistical analysis, but also a machine learning technique known as Gradient Boosting.

The ultimate goal of the study is to present an informed impression of the integration of various Semantic Web technologies in art and culture-related online media and assess both the perceived importance of these technologies and, to an extent, a quantitative measure of that importance in relation to a website's popularity.

2. Theoretical Background

More than two decades have passed since Tim Berners-Lee and his team envisioned the Semantic Web, a form of web content that would be readable and, thus, understandable and comprehensible by machines [1]. One of Semantic Web's most important properties is this ability to give more valuable information by automatically searching the meaning structure of web content [5]. In essence, it would provide structure to the anarchic structure of the Web in conjunction with the focus on human-centered computing as presented by Michael Dertouzos [6], an early proponent of the Semantic Web. The Semantic Web, alongside other web paradigms such as the Social Web which involves the evolution of social media, the 3D Web which encompasses virtual reality experiences in the Web and the Media Centric Web which focuses on the transmediative nature of the Web, is a key consisting element of Web 3.0 [7]. Furthermore, new technologies such as Artificial Intelligence and the blockchain also enhance and improve aspects of the Web, aiming to achieve different goals such as decentralization, connectivity and ubiquity. Web 3.0 and especially the Semantic Web, which is the focus of this study, seem to find their way in multiple thematic fields in the Web, such as news outlets or art and culture-related websites. The rise of increasingly technologically advanced forms of journalism dispels any questions about the technical optimization of contemporary journalism [3].

Despite the fact that now there is more pluralism in terms of websites than ever and more opinions can be heard, the sheer reality is that the majority of this information is left unstructured [3]. Semantic Web solutions could be valuable for journalistic research. The Web offered journalists the plurality that was missing but as a result this led to a more time-consuming process where journalists need to navigate through all the available data and sources and filter the information they are accessing manually [4]. Semantic Web technologies could automatically read, comprehend and include or exclude the useful information and even improve it by reproducing it with added enhancements such as accessibility features [8], or even advanced user personalization features [9].

Heravi and McGinnis [4] note that a combination of technologies will be necessary to provide a Social Semantic Journalism Framework. These technologies would undoubtedly collaborate with each other and serve as inputs and outputs for one another, establishing a procedure capable of addressing the issue that social media poses to journalists and editors as they attempt to determine what is noteworthy in user-generated content.

Another field that could benefit from Semantic Web solutions is that of art and cultural heritage in general. Cultural heritage can be defined as a kind of inheritance to be preserved and passed down to future generations. It is also linked to group identity and is both a symbol of and an essential ingredient in the building of that group's identity [10]. Cultural heritage is vital to understanding earlier generations and the origins of humanity. The Web has enabled local, national and global publishing, explanation and debate.

More and more museums, galleries and art-related institutions are transferring part or all of their collections into the digital space. The quality of a museum's Web presence on the Web can lead, not only to increased website visits, but also to increased physical visitors [11]. However, cultural heritage resources are vast and diverse. They comprise data or information that is highly structured, very unstructured or semi-structured and derived from both authorized and unauthorized sources, and also include multimedia data such as text, images, audio and video data [2]. In order to accommodate the users' needs, many usability-related features need to be implemented [12], but a usability-oriented approach is not the only approach that can help scientists, companies, and schools better understand cultural data. In the world of the Web, the data architecture of digital museum databases is quite diverse and calls for advanced mapping and vocabulary integration. This does

not come as a surprise as libraries, museums and galleries have always had extended and heterogeneous databases in the physical world as well. Several efforts have been conducted in recent years to digitize cultural heritage assets using Semantic Technologies such as RDF and OWL. There are numerous digital collections and applications available today that provide immediate access to cultural heritage content [13]. Additionally, cultural heritage is moving into new media of linear and non-linear storytelling, using audiovisual hypermedia assets and efforts are being made to provide enhanced semantic interaction, in order to transition such content into the Semantic Web [14].

Since the Web is a space full of available information, it goes without saying that someone can find available many sources related to art and culture that do not belong to their organizations but to websites, blogs or platforms focusing on arts and culture. In fact, art or cultural journalism is a distinctive field of journalism. Janssen [15] in his study about coverage of the arts in Dutch newspapers between 1965–1990, divided it in three levels: The first level regards the general newspapers' general portrayal of art, for example, the amount of space devoted to the arts in comparison to other topics. The second level examines disparities in the amount of focus provided to various creative forms or genres by contrasting, for example, classical and rock music coverage. The third level deals with the coverage that artifacts belonging to a certain artistic genre or subfield receive, for instance, the critical response to freshly released films. There was also a classification between cultural artifacts. The first level concerns the standing of the arts in respect to other (cultural) domains; the second level concerns the hierarchical relationships between art forms or genres; and the third level concerns the ranking of works and producers within a particular artistic domain.

Towards realizing their vision for the Semantic Web, the Semantic Web initiative of the World Wide Web Consortium (W3C) has developed a set of standards and tools to support this. Their early work resulted in two significant proposals: the Resource Description Framework Model and Syntax Specification and the Resource Description Framework Schema Specification. The W3C consisted of two primary working groups, the RDF Core Working Group and the Web Ontology Working Group, both of which issued significant sets of recommendations [16]. Since its inception, the Semantic Web has been evolving a layered architecture. Although there have been many variations since, its various components are:

- Unicode and URIs: Unicode as the computer character representation standard, and URIs, as the standard for identifying and locating resources (such as Web pages), offer a foundation for representing characters used in the majority of the world's languages and identifying resources.
- XML: XML and its relevant standards, such as and Schemas and Namespaces, are widely used for data organization on the Web, but they do not transmit the meaning of the data.
- Resource Description Framework: RDF is a basic information (metadata) representation framework that utilizes URIs to identify Web-based resources and a graph model to describe resource relationships. RDF lays the foundation for publishing and linking data. There are a number of syntactic representations available, including a standard XML format.
- RDF Schema: a simple type modelling language for describing classes of resources and properties between them in the basic RDF model. It provides a simple reasoning framework for inferring types of resources.
- Ontologies: a richer language capable of expressing more complicated constraints on the types and attributes of resources.
- Logic and Proof: an (automated) reasoning system built on top of the ontology structure with the purpose of inferring new relationships. Therefore, a software agent can deduce whether a certain resource fits its needs by utilizing such a system (and vice versa).

- Trust: This component is the final layer of the stack, and it refers to the issues of trust and trustworthiness of the information [16]. There are two main approaches regarding trust, the one is based on policy and the second on reputation [17]. Nowadays technology, trust and proof are regarded as the most emerging research areas of the Semantic Web [17].

Semantic Web technologies are regarded as an approach to manage knowledge by utilizing ontologies and semantic web standards, allow individuals to establish data repositories on the Web, create vocabularies, and write rules for data processing. Linked data are assisted by technologies such as RDF, SPARQL, OWL and SKOS [18]. Additionally, a very classic but perhaps outdated framework is the Ranking Semantic Search framework (RSS). This framework enables ranking of the search results on the Semantic Web and, through the use of novel ranking strategies, avoids returning disordered search results [19].

Semantic Web technologies can be applied to a website so it will be more easily readable and accessible by search engines for better Search Engine Optimization (SEO). For instance, website owners or content managers can enhance their text descriptions with semantic annotations and check if this leads to a more satisfying user experience. Towards this end, Necula et al. [20] investigated e-commerce websites and whether there is a correlation between the enhancement of product text descriptions with semantic annotations and the perceived consumers' satisfaction. Their study concluded that the inclusion of semantic web elements in the products descriptions is important for a more pleasant customer experience. In fact, one of the most interesting findings was that the consumer regards knowledge graphs as having high significance in an e-commerce website [20].

A way to add information that is machine-understandable to Web pages that is processed by the major search engines to improve search performance is schema.org [21]. Schema.org's wide adoption is related to its promotion by major search engines as a standard for marking up structured data in HTML web pages [22]. This adoption addresses a fundamental issue for the Web, by making it easy to annotate data within websites, at least for the most common types of Web content [23].

Although using Semantic Web technologies will lead to a more pleasant and usable user experience, it is not certain that this automatically means an improvement in terms of popularity. It is a fact that SEO techniques are used in order to improve a website searchability and consequently popularity, but it is not certain that utilizing Semantic Web technologies will automatically result in increased popularity. This is what this research tries to shed light on.

3. Methodology

3.1. Relevant Website Discovery

In order to gain as much information as possible concerning the level of Semantic Web integration in art and culture-related online media, collecting a big sample of appropriate websites was an essential requirement. Identifying such websites was a complex process involving both automated procedures and human expert input, in order to achieve the best results. The study's website discovery process did not attempt to genuinely discover any and all existing appropriate websites, but instead focused on collecting a sufficiently large sample.

3.1.1. Automated Sampling

The first step in this process was to acquire an up-to-date list of websites belonging to the generic Top Level Domains (gTLDs) that any person or entity is permitted to register, which are the .com, .net and .org gTLDs. The rest of original gTLDs (.int, .edu, .gov and .mil) were excluded since the study's main focus was on private activity, both commercial and non-profit. Such a list was acquired through Common Crawl a "non-profit organization dedicated to providing a copy of the internet to internet researchers, companies and individuals" [24]. The acquired list concerned websites that were indexed in October 2021 thus making it appropriately relevant. In order to pinpoint websites that offered what the

study required, a series of keywords were used. These keywords were "art", "media", "entertain" as short for entertainment and "cult" as short for culture. This starting point of relevant website discovery procured 449,063 Second Level Domains (SLDs), as seen in Table 1.

Table 1. Initially collected SLD quantities.

gTLD	Art	Media	Entertain	Cult
.com	326,153	49,964	7023	10,578
.net	17,577	3195	366	663
.org	28,413	2574	125	2432
Totals	372,143	55,733	7514	13,673

In order to accomplish the above, an automated process was created. This process, which was developed in PHP, used Common Crawl's API to receive a list of domains page by page. Then it proceeded to filter out all subdomains and sub-folders and to check each domain name for the specified keywords. For domains that were available through both HTTP and the more secure HTTPS the non-secure version was filtered out. For domains available both with the www subdomain and without, the ones with www were excluded. If no secure version was available, the non-secure one was kept. The same principle was applied with regards to the www subdomain. This procedure's flowchart can be seen in Figure 1.

Since the websites would be required to be evaluated on their content in order to establish that they are indeed relevant to art or culture and include what may be considered reportorial content, any websites that did not support the English language were excluded. Dual language websites were accepted as long as English was the primary language. This decision was largely influenced by the fact that the English language is in a position of dominance in the Web compared to other languages [25,26]. This is directly related to the "digital divide"—the inequality present in the information society which stems from the difficulty of Internet access for a significant portion of humanity [26,27]. The language of a website was identified based on Common Crawl Index's language parameter. This reduced the number of potentially useful websites to 252,105 sites. Consequently, a number of these websites were filtered out based on the presence in their sTLD of irrelevant words that share part of them with the aforementioned keywords (i.e., earth, heart, quarter, smart, chart, part, etc.)

The next step in narrowing down the number of websites that had to be evaluated was identifying their popularity. Even though the World Wide Web is full of interesting art blogs, cultural publications, artist collectives and more, it is expected that the most popular websites are the ones that have the most impact and the ones worth focusing on. In order to assess each site's popularity, Alexa's Web Information Service was used. Alexa Internet is an Amazon.com company providing insight on site popularity for more than 25 years [28]. The Web Information Service provides information about web sites through the use of a Web services API. Part of this information is the rank of almost 10 million websites based on traffic in the last 90 days. An automated process that used Alexa's Web Information Service through API requests was implemented in order to gather information for all the websites in our database. After eliminating all low-traffic websites, a total of 16,616 websites remained.

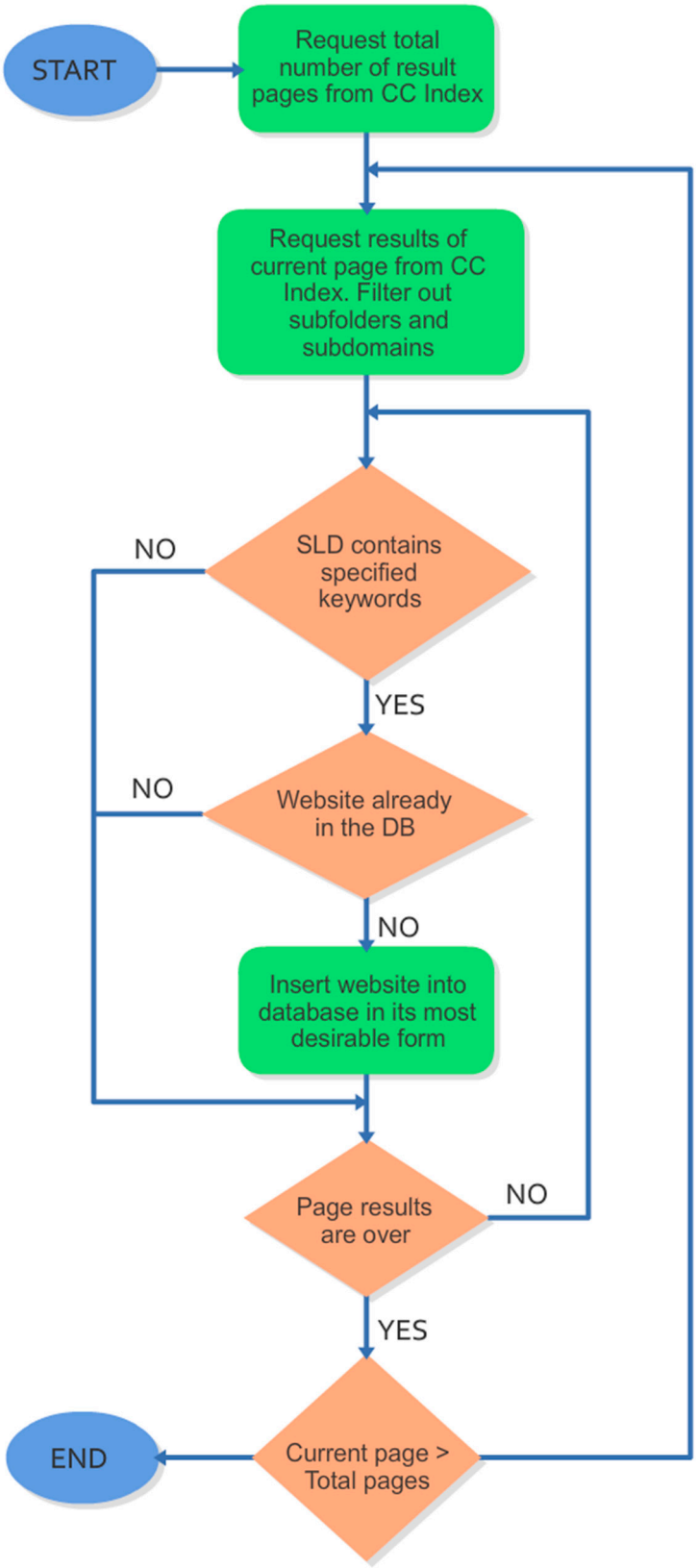


Figure 1. Flow chart of the website discovery crawler.

3.1.2. Expert Screening

In order to identify which of the remaining websites actually corresponded with the researcher's intends the rest of the screening was accomplished through manually visiting and browsing through the websites. This process was conducted by the members of the research team themselves. The Team consists of multiple experts with years of accumulated experience in studying the Web presence of Audiovisual Arts and Cultural Heritage.

In an effort to accommodate this intense and time-consuming process, a small-scale Web application was designed and implemented. Its purpose was to present the researchers with a number of potential websites alongside some information about them like their gTLD and the keyword that their SLDs contained. The members of the research team would then visit the website and after their audit they could choose to evaluate it by selecting the appropriate of three color-coded options:

- Red indicated that a website that had no relevance to the field of art and culture.
- Yellow indicated that a website was art or culture related but had limited reportorial content.
- Green indicated that a website was not only art or culture-related but also contained a fair amount of reportorial content.

The two main criteria for this evaluation were each website's relevance to the fields of art or cultural heritage and whether the website's content was even partially of a journalistic or reportorial nature. Websites that contained information about works of art, artists and their past and current projects, local or international culture and cultural heritage, cultural artifacts or places, historical or academic analysis of artworks, and so on, were considered by the researchers relevant to the fields of art or culture. Such websites included, but were not limited to, websites of museums, galleries, collections, art schools and colleges, artist portfolios, organizations or societies promoting art and culture, news outlets covering relevant matters, artist agencies, art vendors and more. Any non-relevant websites were marked as "Red" and excluded from the study. The researchers also searched for reportorial content inside each website such as articles, blog entries, opinion pieces, artwork analyses, news regarding events or exhibitions, artwork reviews, artist interviews, historical retrospects, current artistic event discussion and more. Websites that exhibited a fair amount of such reportorial content were evaluated as "Green" while those that were relevant to art but exhibited extremely limited or no reportorial content were evaluated as "Yellow".

Figure 2 presents a screenshot of the Web application during its use. The interface also presents the total number of evaluations required, as well as the current number of evaluations completed by this researcher. Additionally, it presents a small preview of how the evaluation process is shaping up by indicating how many websites have been so far evaluated in each category.

Out of a total of 16,616 websites, 12,874 were evaluated as Red, 2653 were evaluated as Yellow and 1089 were evaluated as Green. In addition, the researchers were encouraged to suggest additional websites that they knew fit the criteria and were not discovered by the automated process. This led to an additional 35 websites that were added to the pool and evaluated as Green, bringing the total number of evaluated websites to 16,651 and the total value of Green websites to 1124.

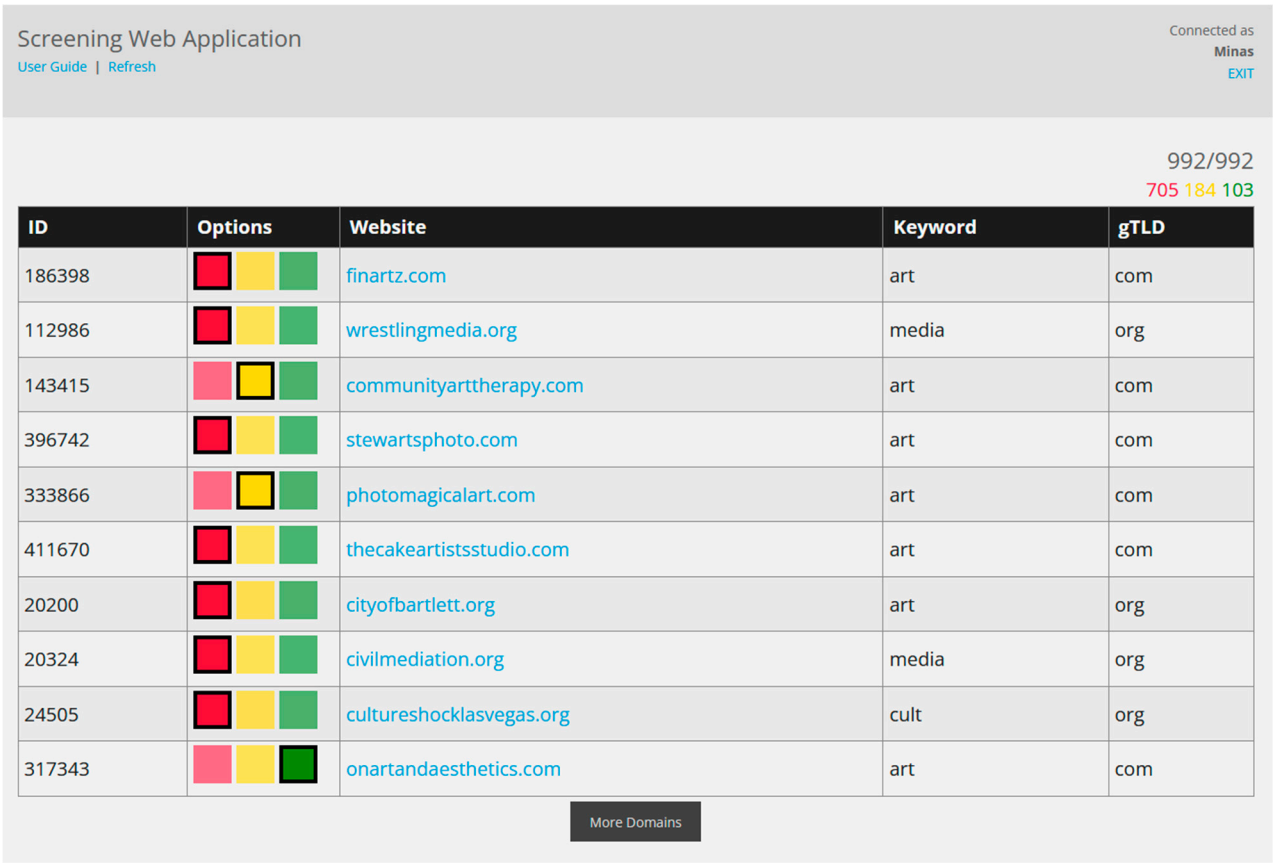


Figure 2. Interface of the Web application created to facilitate the manual screening process.

3.2. Collecting Information

The next step of the study involved investigating the relevant websites in order to collect information regarding which Semantic Web technologies were integrated in each website and, where possible, to what extent. As a means of accomplishing this, an automated procedure was developed and implemented in PHP making use of the cURL Library, a library created by Daniel Stenbergand to allow for connectivity and communication with different servers through various protocols [29], and the DOM PHP extension. All websites evaluated as Yellow and Green were deemed important for this step, since a website is also an information outlet in and of itself. As a result, 3777 websites were investigated.

This procedure connected to the websites' homepage and identified all internal links presented there. It then proceeded to "crawl" through these links and attempted to detect the use of various specific methods that had as a goal to assist with each website's machine-readability. For every website a maximum of 80 pages, including the homepage, were crawled, in an effort to avoid spending an overly extended time in a single website. This number of pages was deemed by the researchers capable of providing a comprehensive impression of the extent of integration of Semantic Web technologies. After identifying these technologies, the crawler also attempted to extract metrics on their usage in manners that will be further elaborated upon below. Out of the 3777 websites, 3632 were successfully crawled. The unsuccessful attempts included websites that denied automated indexing through a robots.txt file or where the crawler encountered various technical difficulties. Figure 3 presents the flow chart of the crawler.

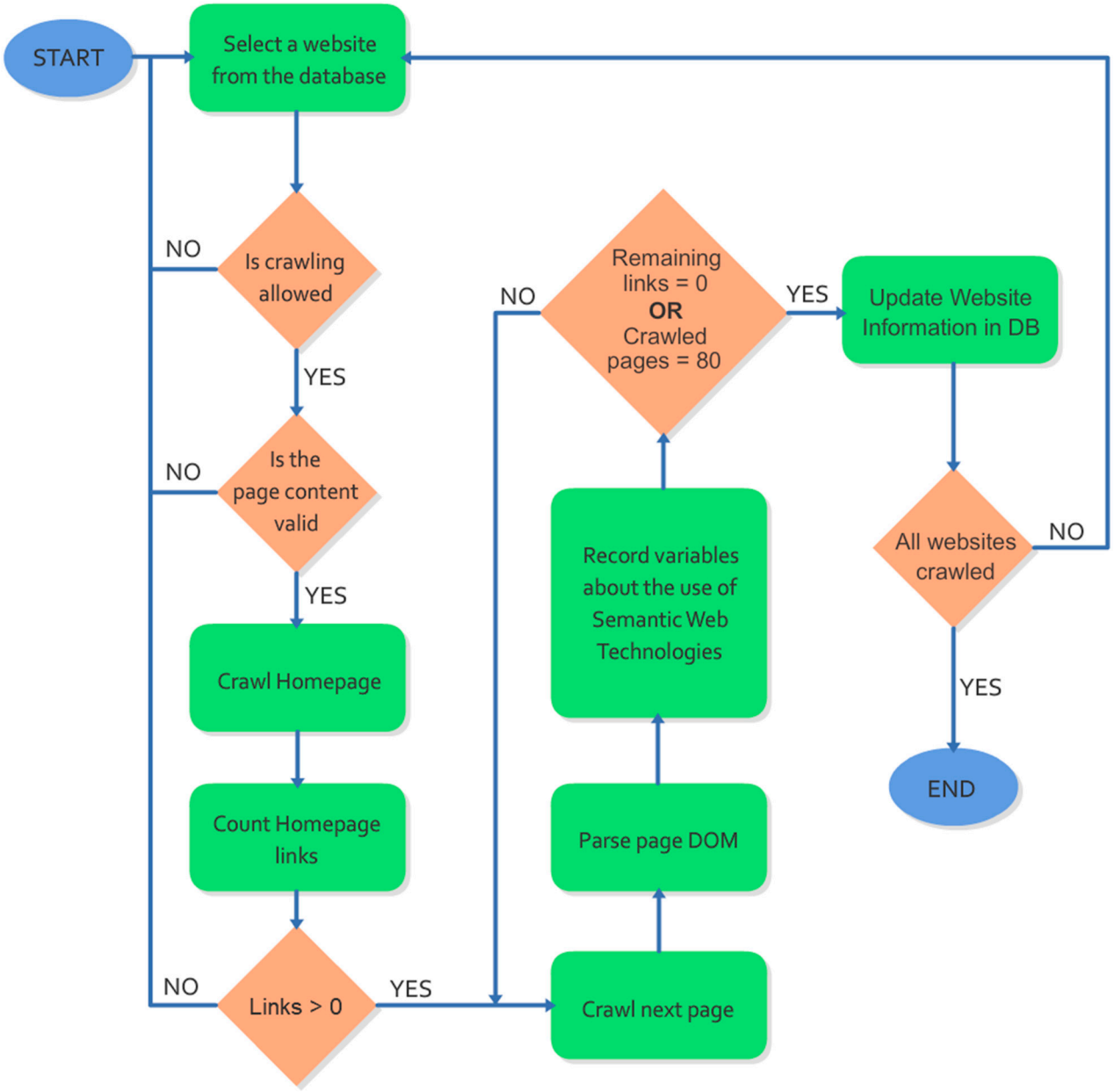


Figure 3. Flowchart of the Semantic Web technologies detection crawler.

The Semantic Web technologies investigated were:

- The use of RSS feeds;
- The use of HTML5 Semantic Elements;
- The use of the Open Graph protocol;
- The use of Twitter Cards markup;
- The use of schema.org schemas;
- The use of Microformats metadata format.

These different methods of creating a more machine-readable website are detailed below.

3.2.1. RSS Feeds

RSS (RDF Site Summary) is a format that allows the creation of Web feeds [30] that can be used to allow applications to access a website's information. It is one of the earliest attempts at Web syndication and through its popularity in the 2000s it stood in the forefront

of creating a machine-readable Web. The study's algorithm detected how many unique RSS feeds were available in each individual website (variable _rss_feeds).

3.2.2. HTML Semantic Elements

The use of new Semantic Elements was introduced in HTML5 in an effort to help define the content of various structural elements of the Document Object Model (DOM) not only to the developer but also to the browser [31]. These are specifically the elements <article>, <aside>, <details>, <figcaption>, <figure>, <footer>, <header>, <main>, <mark>, <nav>, <section>, <summary> and <time>. The study's crawling algorithm located such elements in a Web page's structure and counted how many pages of each website included these elements (variable _html). Additionally, it monitored how many different such elements were used for each individual website (variable _html_variety).

3.2.3. Open Graph

The Open Graph protocol contains metadata that help create a rich object regarding each Web page for the purpose of displaying it in a social graph [32]. The protocol follows a method compatible with W3C's RDFa (Resource Description Framework in attributes) recommendation. The most basic metadata element of the protocol is the og:title element which contains a title for the Web page as it would appear on the graph. The study's algorithm detected how many pages of each website included an og:title element (variable _og). Additionally, it monitored what percentage of these titles where unique to a single specific page and not a reused generic title (variable _og_variety).

3.2.4. Twitter Cards

Twitter Cards use a markup system to create a rich object specifically for the Twitter social media platform [33]. Similarly to the Open Graph system, it complies with RDFa syntax. The Summary card is a twitter card which creates a short summary of the specific Web page. The title of that summary can be located in the twitter:title metadata element. The study's algorithm detected how many pages of each website included a twitter:title element (variable _twitter). Additionally, it monitored what percentage of these titles where unique to a single specific page and not a reused generic title (variable _twitter_variety).

3.2.5. Schema.org Schemas

Schema.org is a community-driven collection of structured data schemas [22] for use on the Internet, founded by Google, Microsoft, Yahoo and Yandex. Its purpose is to make it easier on website developers to integrate machine-readable data in their websites. The data can be conveyed using different encodings such as RDFa, Microdata or JSON-LD. The study's algorithm detected how many pages of each website included a schema.org element in any of these three different methods of encoding (variable _schemaorg).

3.2.6. Microformats

Microformats is a set of data formats that can be used to convey machine-readable data. The various data formats are explicitly declared through the use of HTML Classes [34]. Multiple such Microformats are available in order to semantically mark a variety of information. The study's algorithm detected how many subpages of each website included one of the following classes indicated usage of a microformat: "adr", "geo", "hAtom", "haudio", "vEvent", "vcard", "hlisting", "hmedia", "hnews", "hproduct", "hrecipe", "hResume", "hreview", "hslice", "xfolkentry" and "xoxo" (variable _microformats). Additionally, it monitored how many different such classes were used for each individual website (variable _microformats_variety).

In addition to the above technologies, the crawling algorithm developed in this study kept a record on how many pages of each website were crawled (variable _pages_crawled) as well as any other json + app formats that might appear in a page that might be worth

investigating (variable _other). A comprehensive table of all variables recorded by the study's algorithm is presented in Table 2.

Table 2. List of SWT related variables recorded by the crawler algorithm.

Variable	Short Description
_pages_crawled	Number of pages crawled
_rss_feeds	Number of unique RSS feed links
_html	Number of pages with HTML5 Semantic Elements
_html_variety	% of HTML5 Semantic Elements used
_og	Number of pages with Open Graph Metadata Elements
_og_variety	% of og:title values that are unique
_twitter	Number of pages with Twitter Summary Card Metadata Elements
_twitter_variety	% of twitter:title values that are unique
_schema_org	Number of pages with schema.org structured data
_microformats	Number of pages with Microformats data formats
_microformats_variety	% of Microformats used
_other	Number of pages with other json data

3.3. Evaluating Semantic Web Technologies Integration

The multitude of measured quantitative variables that were collected during the website crawling process are all indicators of a website's adherence to Semantic Web standards. They can be used to get a glimpse of how committed each website is to making its information machine-readable. As part of the effort of documenting this commitment the researchers have created a 5-star rating system that can translate the measurements in an easy-to-read comprehensive value dubbed "Semantic Web Technologies Integration Rating" or SWTI rating.

The rating system focuses on which elements the researchers consider most important through their expertise in the field of integration of Semantic Web technologies.

The usage of structured data is the first and most important aspect of such an integration. Schema.org is supported by multiple colossi of the Web such as Google and Microsoft and Microformats has a long history of effort in promoting machine-readability. Hence, one star is rewarded for attempting use of these technologies at least to some extent, with a second star being rewarded to websites that have a more extensive integration.

The creation of rich objects for social media may stem from a different motivation but nonetheless it is a major contributing factor in the machine-readability of modern websites. As such, one star is awarded for the implementation of at least one such method, either Open Graph or Twitter Cards. An additional half star is awarded if the implementation focuses in providing unique information for each different page of a website, as dictated by usage guidelines.

The use of HTML5 semantic elements promotes a content-based structure of a website's DOM at least to some extent and so it is rewarded with half a star. Additionally, when the website uses multiple different such elements it becomes an indicator of a quality implementation and as such it is rewarded with another half star.

Finally, providing an RSS feed has been a popular practice for more than 20 years and it is a good first step in assisting with machine-readability. Since RSS popularity is declining, its use awards half a star.

The scoring system is presented in detail in Table 3.

Table 3. SWTI rating system.

Condition	Stars Awarded
Website using at least some schema.org or Microformats structured data.	1
Website using Schema.org or Microformats structured data in at least 50% of its crawled pages	1
Website using Open Graph or Twitter Cards to provide at least one rich object for social media	1
Website using unique Open Graph or Twitter Card titles for over 50% of its rich objects	0.5
Website using at least one HTML5 Semantic Element	0.5
Website using at least 50% of the different HTML5 Semantic Elements	0.5
Website providing at least one RSS feed	0.5
Total	5

4. Results

4.1. Statistics and Analysis

Table 4 depicts a sample of the first ten entries of our data formation. The first column shows the websites, the second column presents the ranking of a website based on its popularity according to Alexa Internet (_alexa_rank). Columns 3 to 7 contain scores for each of the four major Semantic Web technologies derived by dividing the pages that used each technology as shown in Section 3.2 (variables _rss_feeds, _html, _og, _twitter, _schema_org) by the total number of pages crawled (variable _pages_crawled) thus creating the new variables (_rss_score, _html_score, _og_score, _twitter_score, _schema_score). Columns from 8 to 10 contain the variables _html_variety, _og_variety and _twitter_variety. The last column contains the rating for each site based on the SWTI rating system detailed in Section 3.3 (variable _swti). The variables _microformats, _microformats_variety and _other, which identified usage of microformats or other json data in each web page were omitted from further statistical analysis because the percentage of websites with findings in these metrics was below 1%.

Table 4. Data formation sample.

Domain	Alexa Rank	RSS Score	Html Score	OG Score	Twitter Score	Schema Score	Html var%	OG var%	Twitter var%	SWTI Rating
03mediainc.com	228,588	60.00	95.00	95.00	0.00	95.00	77	97	0	5.00
10xplusmedia.com	269,411	38.46	96.15	96.15	0.00	96.15	31	96	0	4.50
13artists.com	4,084,650	27.27	90.91	90.91	0.00	81.82	23	100	0	4.50
1531entertainment.com	1,278,072	62.50	100.00	100.00	100.00	0.00	77	100	70	3.00
1913mediagroup.com	4,799,977	60.00	100.00	0.00	0.00	0.00	31	0	0	1.00
1artworks.com	4,141,675	65.22	91.30	91.30	86.96	0.00	31	100	5	2.50
1atbatmedia.com	2,355,870	47.06	100.00	100.00	0.00	100.00	54	94	0	5.00
1media-en.com	2,216,622	101.25	71.25	98.75	0.00	98.75	46	97	0	4.50
03mediainc.com	228,588	30.00	17.50	85.00	0.00	85.00	31	100	0	4.50
10xplusmedia.com	269,411	0.00	100.00	96.25	0.00	100.00	23	99	0	4.00

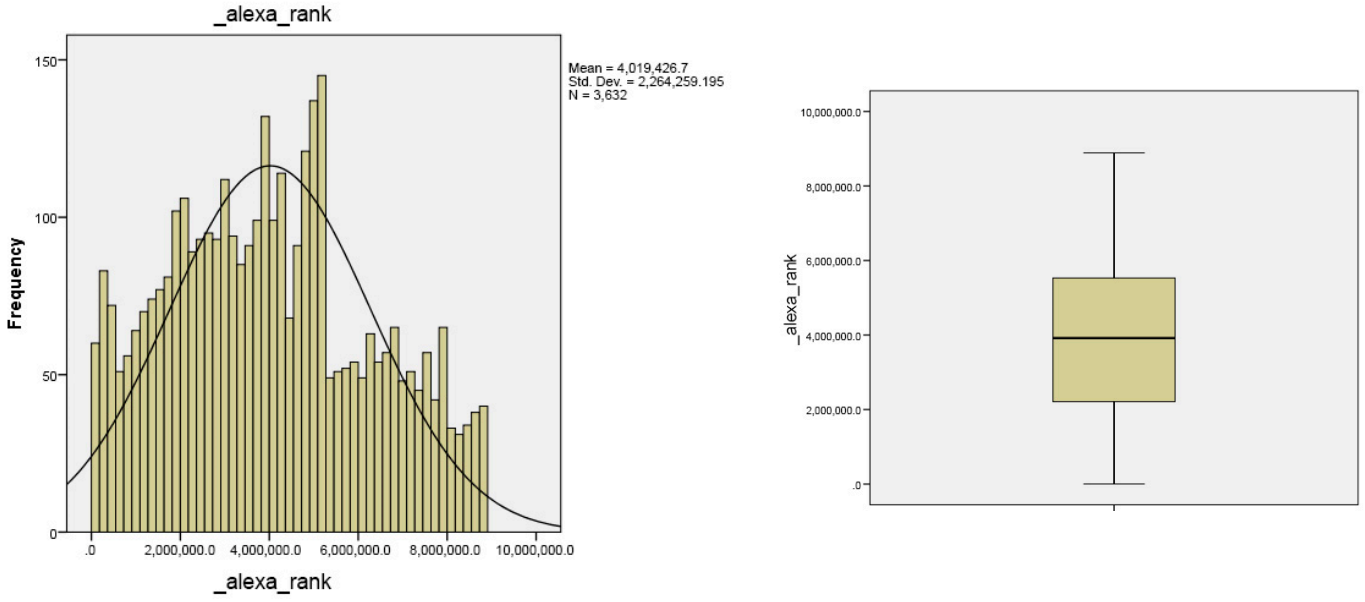
Table 5 depicts the descriptive statistics for each variable and Table 6 depicts the frequency related statistics. In Figure 4 the histogram and boxplot of the _alexa_rank variable are presented, depicting the distribution and dispersion of the variable. The histogram and boxplot, the distribution and dispersion of the other sample values are plotted for each of the variables and presented in Appendix A.

Table 5. Descriptive statistics.

	N	Minimum	Maximum	Mean	Std. Deviation
_alexa_rank	3632	3736.0	8,887,606.0	4,019,426.703	2,264,259.1955
_rss_score	3632	0.0	300.0	20.325	33.6279
_html_score	3632	0.0	100.0	77.128	35.1675
_og_score	3632	0.0	100.0	65.814	41.7498
_twitter_score	3632	0.0	100.0	31.371	43.5979
_schema_score	3632	0.0	100.0	42.751	44.2935
_swti	3632	0.0	5.0	2.968	1.5851
_html_variety	3632	0.0	92.0	36.697	21.3882
_og_variety	3632	0.0	100.0	70.367	42.8166
_twitter_variety	3632	0.0	100.0	33.598	44.1914
Valid N (listwise)	3632				

Table 6. Frequencies.

		_alexa _rank	_rss _score	_html _score	_og _score	_twitter _sctore	_schema _score	_html _variety	_og _variety	_twitter _variety	_swti
N	**Vald**	3632	3632	3632	3632	3632	3632	3632	3632	3632	3632
	Mis	0	0	0	0	0	0	0	0	0	0
Mean		4,019,426.70	20.325	77.128	65.814	31.371	42.751	36.697	70.367	33.598	2.968
Std. Err. of Mean		37,571.0403	0.5580	0.5835	0.6928	0.7234	0.7350	0.3549	0.7105	0.7333	0.0263
Std. Deviation		2,264,259.19	33.6279	35.1675	41.7498	43.5979	44.2935	21.388	42.8166	44.1914	1.5851
Variance		5.127×10^{12}	1130.837	1236.75	1743.047	1900.77	1961.915	457.45	1833.257	1952.879	2.513
Skewness		0.225	2.702	−1.476	−0.777	0.772	0.226	−0.177	−0.949	0.634	−0.315
Kurtosis		−0.808	11.401	0.534	−1.206	−1.309	−1.810	−0.703	−1.003	−1.501	−1.190
Range		8,883,870.0	300.0	100.0	100.0	100.0	100.0	92.0	100.0	100.0	5.0
Minimum		3736.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Maximum		8,887,606.0	300.0	100.0	100.0	100.0	100.0	92.0	100.0	100.0	5.0
Percent	**25**	2,210,762.75	0.000	75.000	5.212	0.000	0.000	23.000	7.000	0.000	1.500
	50	3,915,896.50	3.101	95.000	90.909	0.000	20.000	38.000	100.000	0.000	3.000
	75	5,530,911.25	28.571	100.000	100.000	90.183	94.118	54.000	100.000	89.000	4.500



(**a**) (**b**)

Figure 4. The histogram (**a**) and boxplot (**b**) of the _alexa_rank variable showing a relatively normal distribution.

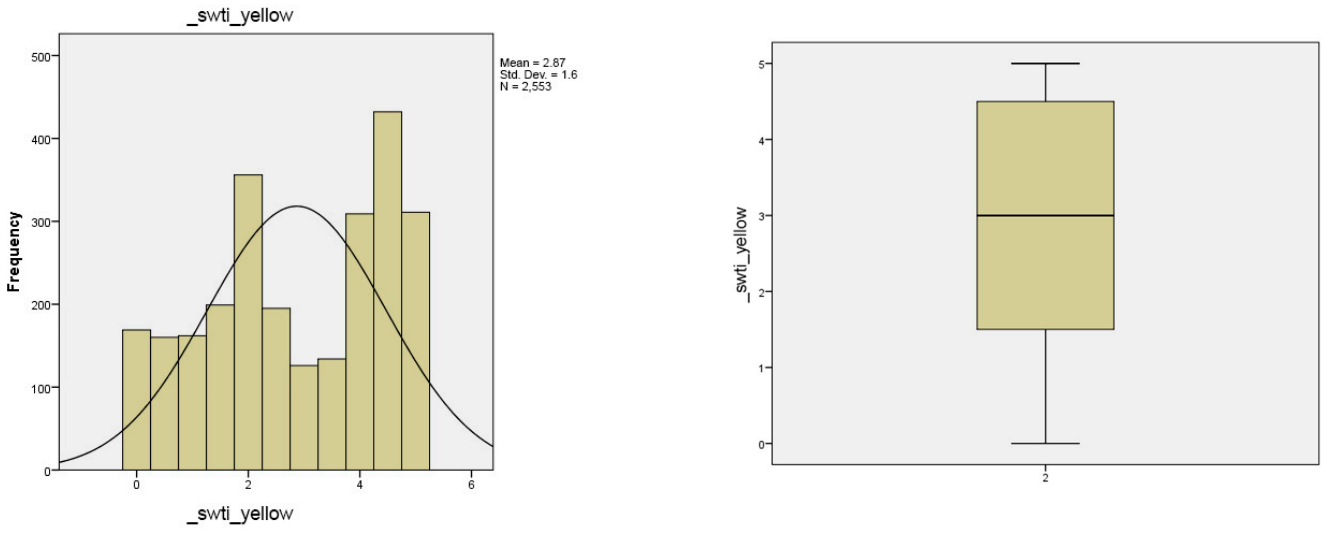
Table 7 depicts the descriptive statistics of the _swti variable for websites that were evaluated as Yellow or Green during the expert screening process described in Section 3.1.2 (variables _swti_yellow, swti_green). Frequency-related statistics for these variables are presented in Table 8 and their histograms and boxplots in Figures 5 and 6.

Table 7. Descriptive statistics for the SWTI of Yellow and Green websites.

	N	Minimum	Maximum	Mean	Std. Deviation
_swti_yellow	2553	0	5	2.87	1.600
_swti_green	1079	0	5	3.20	1.524
Valid N (listwise)	1079				

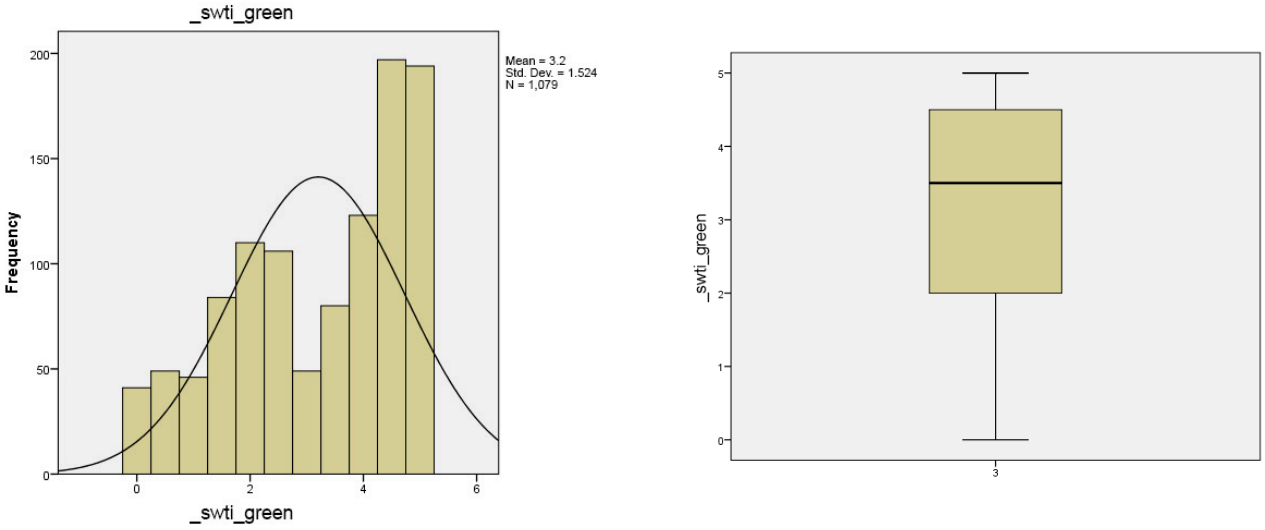
Table 8. Frequencies for the SWTI of Yellow and Green websites.

		_swti_yellow	_swti_green
N	**Valid**	2553	1079
	Missing	1079	2553
Std. Error of Mean		0.032	0.046
Std. Deviation		1.600	1.524
Variance		2.561	2.321
Skewness		−0.244	−0.483
Std. Error of Skewness		0.048	0.074
Kurtosis		−1.245	−0.997
Std. Error of Kurtosis		0.097	0.149
Range		5	5
Minimum		0	0
Maximum		5	5
Percentiles	**25**	1.50	2.00
	50	3.00	3.50
	75	4.50	4.50



(**a**) (**b**)

Figure 5. The histogram (**a**) and boxplot (**b**) of the _swti_yellow variable showing a non-normal distribution.



(**a**) (**b**)

Figure 6. The histogram (**a**) and boxplot (**b**) of the _swti_green variable showing a non-normal distribution.

In order to analyze the interrelation between the independent variables _html_score, og_score, _twitter_score and _schema_score, a Pearson's r criterion was applied [35]. The results are shown in Table 9 where the correlations between the variables are depicted. All the correlations have significance level at 0.000 confirming the statistically significant correlation.

Table 9. Pearson correlations between html_score, og_score, twitter_score and schema_score.

		html_score	og_score	twitter_score	schema_score
html_score	**Pearson correlation**	1	0.393 **	0.316 **	0.349 **
	sig.		0.000	0.000	0.000
	N	3632	3632	3632	3632
og_score	**Pearson correlation**	0.393 **	1	0.482 **	0.422 **
	sig.	0.000		0.000	0.000
	N	3632	3632	3632	3632
twitter_score	**Pearson correlation**	0.316 **	0.482 **	1	0.099 **
	sig.	0.000	0.000		0.000
	N	3632	3632	3632	3632
schema_score	**Pearson correlation**	0.349 **	0.422 **	0.099 **	1
	sig.	0.000	0.000	0.000	
	N	3632	3632	3632	3632

** Correlation is significant at the 0.01 level.

The Pearson's r coefficients range from 0.316 (the weaker positive correlation between html_score and twitter score) to 0.482 (the stronger positive correlation between og_score and twitter score).

In an effort to examine the interrelationship between the collected Semantic Web metrics and a websites popularity the various websites were ranked according to their measured SWTI rating (variable _swti_rank) and the Spearman's rank correlation coefficient was calculated. The results are presented in Table 10.

Results of the Spearman correlation indicated that there is a significant very small positive relationship between _swti_rank and _alexa_rank and Y, ($r(3630) = 0.0683$, $p < 0.001$). This correlation despite being very small prompted the researchers to further investigate the interrelationship between SW integration and popularity using a gradient boosting analysis which included every metric collected by the crawling algorithm.

Table 10. Spearman correlation between _swti_rank and _alexa_rank.

Parameter	Value
Spearman correlation coefficient (r)	0.06835
***p*-value**	0.0000375
Covariance	161,169,282.6
Sample size (n)	3632
Statistic	4.1275

4.2. Gradient Boosting Analysis Using XGBoost

After samples have been collected, the XGBoost models are built using a grid search among the parameter space. XGBoost (eXtreme Gradient Boosting) is a fast implementation of gradient boosting [36]. It is a scalable end-to-end tree boosting system that has been widely used and achieves state-of-the-art classification and regression performance [37]. It can improve in the reduction of overfitting, the parallelization of tree construction, and the acceleration of execution. It is an ensemble of regression trees known as CART [38]. The prediction score is calculated by adding all of the trees together, as indicated in the following equation,

$$\hat{Y} = \sum_{m=1}^{M} f_m(X) \tag{1}$$

where M is the number of trees and f_m is the mindependent CART tree. In contrast to Friedman's [39] original gradient boosting architecture, XGBoost adds a regularized objective to the loss function. The regularized objective for the mth iteration optimization is provided by

$$L^m = \sum_{i=1}^{n} l(y_i, \hat{y_l}^m) + \sum_{j=1}^{m} \Omega(f_j) \tag{2}$$

where n denotes the number of samples, l denotes the differentiable loss function that quantifies the difference between the predicted $\hat{y_i}^m$ and the target y_i and Ω denotes the regularization term

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{k=1}^{T} {w_k}^2 \tag{3}$$

where T is the number of nodes and w denotes each node's weight. The regularization degree is controlled by two constants, γ and λ. Furthermore, taking into account that for the mth iteration the following relation holds,

$$\hat{y_l}^m = \hat{y_l}^{m-1} + f_m(x_i) \tag{4}$$

we can recast Equation (2) as,

$$L^m = \sum_{i=1}^{n} \left[l\left(y_i, {y_i}^{m-1}\right) + g_i f_m(x_i) + \frac{1}{2} h_i {f_m}^2(x_i) \right] + \Omega(f_m) \tag{5}$$

where we introduced the operators, $g_i = \partial_{\hat{y}^{m-1}} l\left(y_i, \hat{y}^{m-1}\right)$ and $h_i = \partial^2_{\hat{y}^{m-1}} l\left(y_i, \hat{y}^{m-1}\right)$, which are the loss function's first and second-order derivatives, respectively.

XGBoost makes the gradient converge quicker and more accurately than existing gradient boosting frameworks by using the second-order Taylor expansion for the loss function [36]. It also unifies the generation of the loss function's derivative. Furthermore, adding the regularization term XGBoost to the target function balances the target function's decrease, reduces the model's complexity, and successfully resolves overfitting [36].

Furthermore, XGBoost can use the weight to determine the importance of a feature. The number of times a feature is utilized to partition the data across all trees is the weight in XGBoost [36], and is given by the equation

$$IMP^F = \sum_{m=1}^{M} \sum_{l=1}^{L-1} I\left(F^l{}_m,\ F\right) I\left(F^l{}_m,\ F\right) \tag{6}$$

with the boundary conditions, $IMP^F = 1$, if $F^l{}_m == F$, else $IMP^F = 0$. M is the number of trees or iterations, L denotes the number of nodes in the mth tree, $L-1$ denotes the tree's non-leaf nodes, $F^l{}_m$ stands for the corresponding feature to node l, and $I()$ denotes the indicator function.

The Alexa ranking for the websites under investigation is used as the outcome of the fitted model. The features collected by the crawling mechanism are used as the predictor variables. Since the main point of the analysis is to identify the most important features related to semantic web technologies with respect to the ranking of a website, we perform a grid search for the parameter space of XGBoost. The Alexa ranking is used to extract four classes using the quartiles with respect to the ranking. This transforms the regression analysis to a multiclass classification problem with four classes available. The first class is for the top 25% of the websites in ranking, and the other three classes are for the intervals [0%, 25%), [25%, 50%) and [50%, 75%] of the remaining websites.

The measure logLoss, or logarithmic loss, penalizes a model's inaccurate classifications. This is particularly useful for multiclass classification, in which the approach assigns a probability to each of the classes for all observations (see, e.g., [40]). As we are not expecting a binary response, the logLoss function was chosen over traditional accuracy measurements. The logLoss function is given by

$$logLoss = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} ln(p_{ij}) \tag{7}$$

where, M is the number of classes, N the number of observations, $y_{ij} = \{0,\ 1\}$ indicates if observation i belongs to class j, an p_{ij} the respective probability.

The number of pages crawled is used to scale the related features extracted. These are all page count features for a respective semantic web technology, and the feature extracted for the rss feeds. In addition, this transformation "scales-out" the number of pages crawled to isolate the effect, and the importance of the semantic web features measured to ranking. In particular, the following variables are transformed by dividing with the number of pages crawled ("_pages_crawled"), "_html", "_og", "_twitter", "_rss_feeds", "_schema_org", "_other", "_microformats".

The parameters of machine learning models have a significant impact on model performance. As a result, in order to create an appropriate XGBoost model, the XGBoost parameters must be tuned. XGBoost has seven key parameters: boosting number (or eta), max depth, min child weight, sub sample, colsample bytree, gamma, and lambda. The number of boosting or iterations is referred to as the boosting number. The greatest depth to which a tree can grow is represented by max depth. A larger max depth indicates a higher degree of fitting, but it also indicates a higher risk of overfitting. The minimum sum of instance weight required in a child is called min child weight. The algorithm will be more conservative if min child weight is set to a large value. The subsample ratio of the training instances is referred to as subsample. Overfitting can be avoided if this option is set correctly. When constructing each tree, colsample bytree refers to the subsample ratio of features. The minimum loss reduction necessary to make a further partition on a tree leaf node is referred to as gamma. The higher the gamma, the more conservative the algorithm is. Lambda represents the L2 regularization term on weights. Additionally, increasing this value causes the model to become more conservative. We perform a grid search using the facilities of the Caret R-package [41]. We search the parameter space with the "grid" method, using 10-fold cross validation for a tuneLength of 30 that specifies the total number of unique combinations using the trainControl and train functions of the Caret package (Kuhn 2008). The optimal values identified are

{eta = 0.3, gamma = 0, min child weight = 5, max depth = 6, subsample = 0.5, colsample_bytree = 0.5, lambda = 0.5}.

The overall statistics are presented in Table 11 and the statistics by class in Table 12.

Table 11. Overall statistics.

Overall Statistics	
Accuracy	0.304
95% CI	(0.2742, 0.335)
***p*-Value**	0.0009703
Kappa	0.0727

Table 12. Statistics by class.

	Class: 1	Class: 2	Class: 3	Class: 4
Sensitivity	0.3259	0.23504	0.4081	0.25110
Specificity	0.7968	0.79228	0.6905	0.79295
Pos Pred Value	0.3443	0.28205	0.3003	0.28788
Neg Pred Value	0.7830	0.74895	0.7818	0.76056
Precision	0.3443	0.28205	0.3003	0.28788
Recall	0.3259	0.23504	0.4081	0.25110
F1	0.3349	0.25641	0.3460	0.26824
Prevalence	0.2467	0.25771	0.2456	0.25000
Detection Rate	0.0804	0.06057	0.1002	0.06278
Detection Prevalence	0.2335	0.21476	0.3337	0.21806
Balanced Accuracy	0.5613	0.51366	0.5493	0.52203

Figure 7 presents the sorted accuracy for each model fit and Figure 8 displays the various variables and their importance.

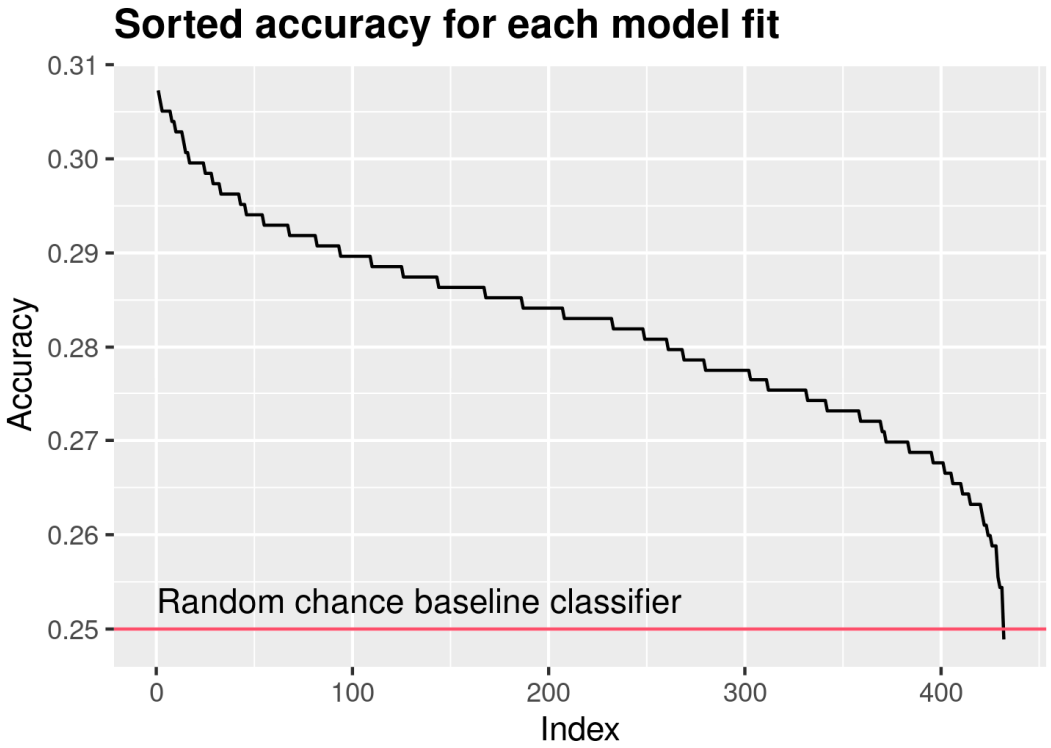


Figure 7. Sorted accuracy for each model fit.

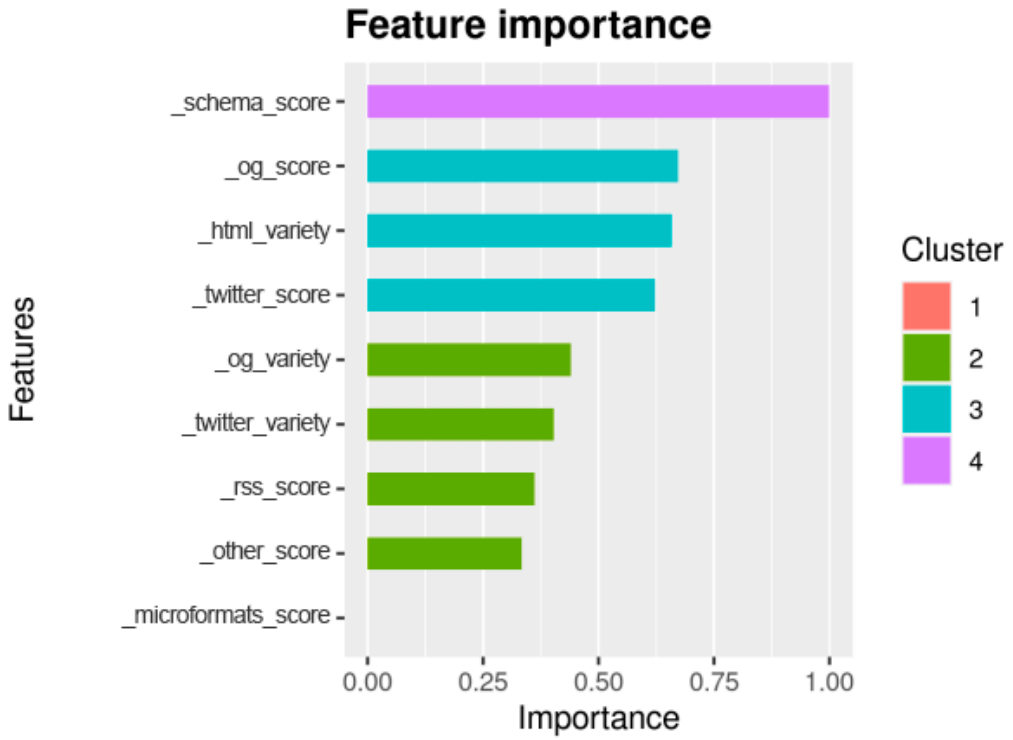


Figure 8. Feature importance.

5. Discussion

The metrics presented in Section 4.1 provide an interesting overview of the landscape of Semantic Web technologies integration in media websites with content relevant to art and cultural heritage. There are some common patterns that seem to emerge while at the same time each variable presents something unique.

The variable _rss_score provides information about the quantity of different RSS feeds found in a single website in relation to the total pages crawled for that website. As seen in the variable's histogram in Figure A1a the relevant majority of websites do not provide RSS feeds at all. That being said, the sites that do provide RSS in total are more than the sites that do not. From the websites that do use RSS as a means to disseminate content most have an _rss_score value from 1–100. This indicates that they provide one or less unique RSS feed per page crawled. With the arithmetic mean being ~20, as seen in Table 5, that would mean that the average website provided 1 RSS feed per 5 pages crawled. This makes sense since usually RSS feeds contain multiple records of content (articles, products, comments, etc.) It was observed that a common practice was to provide some general feeds with multiple records while at the same time providing an additional single record feed in a single article or artwork page. Very few websites seem to provide an abnormally large number of RSS feeds. The vast majority of these websites are sites with very few pages crawled (one or two), which included multiple feeds. These cases account for less than 2% of total websites and might be the result of technical irregularities. In general, RSS usage seems to remain somewhat popular despite the technology being past its prime. A contributor to this might be the fact that many popular Content Management Systems (such as WordPress or Wix) provide RSS feed support out-of-the-box.

The histogram of the variable _html_score in Figure A2a represents a duality: A large number of websites use the HTML Semantic Element tags in all of their pages and another smaller number in none. This is to be expected since adoption of such a technology often happens in a vertical manner throughout all the pages of a website. The mean of the variable is ~77, indicating that usage of at least some HTML Semantic Elements is rather popular. We can obtain more insights regarding these elements through the _html_variety variable. Its histogram in Figure A6a, shows a relatively normal distribution. The peak of

that distribution is at 46% which means that most websites that use these elements use 6 out of 13 identified elements. Although the variety of elements used could be higher, the overall assessment of HTML5 Semantic Elements is encouraging since the technology is both popular and focused in more than a few different elements. Further observation in this area can provide information on the more or least used such elements but it is beyond the scope of this study.

The social media related variables _og_score and _twitter_score have similar histograms, as seen in Figures A3a and A4a, with the bulk of websites either fully committing to the technology or not implementing it at all. This behavior that was also noted in _html_score seems to form a pattern. Open Graph seems to be the more popular of the two with a mean of ~65 vs. one of ~31 as seen in Table 5. This is to be expected since Open Graph is used by multiple social media platforms and messaging systems to create rich previews. Even Twitter itself will create a rich object through open graph if no Twitter card is available. A Twitter card can even indicate that a preview's content can be collected by the appropriate Open Graph meta elements. Looking at the _og_variety and _twitter_variety variables in Figures A7a and A8a, we can note that most websites that implement the technologies also ensure that they provide unique information for each different page of the website. This builds into the already established pattern that when a website developer decides to implement such a technique the implementation is usually comprehensive. Although fewer, there are still cases of websites that provided non-unique titles for the rich object preview.

The structured data related variable _schema_score showed a moderate usage of the schema.org vocabularies throughout the websites included in this study as it was indicated by its mean which is at ~42 as seen in Table 5. In its histogram in Figure A5a we notice the same behavioral pattern as other similar variables. In contrast, the other variables used to identify structured data usage (_microformats, _microformats_variety, _other) all recorded very low usage around 1%. A secondary crawling trying to identify elements with Microformats v2 classes yielded even fewer websites. This indicates that website developer efforts towards implementing structured data is for the time being focusing mainly on schema.org which is founded and supported by major players in the field of SEO and the Web in general.

The Semantic Web Technologies Integration rating as described in Section 3.3 tries to summarize all above metrics in an overall rating (variable _stwi). This variable had a mean of ~2.9 as seen in Table 5 which indicates and above average integration and a standard deviation of ~1.5. In Table 6 the percentile breaking points at 25%, 50% and 75% for this variable are 1.5, 3 and 4.5 which can be interpreted as an indicator of the rating system's quality. In the histogram of the variable, as seen in Figure A9a, we notice two peaks, one around rating value 2 and one around rating value 4.5. This double peak impression can be a result of the behavioral pattern of either implementing a technology fully or not at all that we discerned in the histograms of other individual variables.

As described in Section 3.1.2, the websites were screened by the researchers and split into categories: Red websites, that were outside the study's scope and were not crawled, Yellow, which indicated that a website was art or culture related but had limited reportorial content and Green, which indicated that a website was not only art or culture related but also contained a fair amount of reportorial content. In Section 4 we proceeded to distinguish the information between these two classes, thus creating the variables _swti_yellow and _swti_green. We can see from Table 7 that the SWTI for Green websites has a mean of ~3.2 which is not only greater than that of the yellow websites but also greater than the overall mean of _swti. Additionally, in the histograms of these new variables we notice an overall shift of frequency values towards higher STWI ratings. This is a fair indicator that websites that purposefully provide more journalistic or reportorial contact concerning art and cultural heritage also put more effort into implementing Semantic Web technologies.

Studying the interrelationship between several of the variables that were calculated using metrics from the crawling algorithm described in Section 3.2, there appear to be

multiple moderate positive correlations between them, as seen in Table 9. This is a clear indication that when developers decide to start integrating Semantic Web technologies in their websites, they will often branch to multiple such technologies in order to achieve more comprehensive coverage. The strong correlation between _og_score and _twitter_score is also notable since it demonstrates the importance of multiple technologies when focusing on social media. Developers do not always go for one technology over the other, but they display a notable preference to implement both.

The Spearman correlation analysis between the ranking of the websites based on their SWTI rating and their Alexa ranking indicates a very small, yet significant positive correlation which means that to a small extent, usage of Semantic Web technologies and website popularity are indeed positively related. This can indicate both that websites that are popular are more keen to invest in Semantic Web integration and that Semantic Web integration might actually provide a minimal boost to popularity. To make more of this linear relationship the researchers proceeded to a gradient boosting analysis the results of which were presented in Section 4.2.

The Gradient Boosting analysis was performed as mentioned using the XGBoost algorithm and provided some interesting findings. We can see from the overall statistics presented in Table 11 that overall prediction accuracy surpassed the value of 0.3. Considering the random prediction accuracy for the four defined intervals would be 0.25 there appears to be a small but noticeable increase. The increase's persistence can be observed by the minimum and maximum values of the 95% confidence interval which are both above the baseline of 0.25. Moreover, this increase indicates that, even though Semantic Web integration as measured by this study is not directly correlated with each website's overall popularity, it can still be used to an extent to more accurately predict under which popularity class a website would fall.

By assessing the statistics by class as seen in Table 12, it appears that Class 1, which includes the top 25% of the websites in ranking, displays higher values than the other classes in Positive Prediction Value, Precision and Balanced Accuracy. This might indicate higher credibility of Semantic Web metrics when attempting to predict the popularity of top-ranking websites.

In Figure 8 the features used in the Gradient Boost analysis are presented by order of calculated importance in the accurate prediction of popularity. They are clustered in four groups according to that importance. First and only feature in the most important cluster is the "_schema_org" feature which is an indicator of the percentage of crawled pages that include schema.org structured data. Usage of the schema.org vocabularies is promoted by Google, Microsoft, Yahoo, Yandex and more search engine providers which means that their inclusion to a larger extent, not only provides machine-readable content, but also increases the website's Search Engine Optimization Score (SEO) which in turn influences popularity.

In the second cluster the features _og, _twitter and _html_variety appear. The first two assist with social media integration and thus make a page easier to diffuse through the multiple social media platforms available. The _html_variety feature represents the effort, from a developer's perspective, to enhance a web page's semantic value by using a greater variety of HTML5 semantic elements.

The other features appear to have less importance and are grouped in the remaining clusters. Social media-related rich-data content variety as indicated by "_og_variety" and "_twitter_variety" seems to matter but to a smaller extent. This makes sense if we consider that content can sometimes be accurately described even without much variation in the description itself. The feature "_rss_feeds" which indicates the usage of RSS also plays a more minor role. RSS, though still useful, seems to be of waning importance as a means to convey machine readable information. Additionally, all metrics relating to microformats appear to be irrelevant. This is to be expected judging by how few implementations of this Semantic Web tool were detected during the crawling process.

6. Conclusions

The Semantic Web since its conception has been embraced by people in the fields of art and cultural heritage, because it provides valuable tools for organizing and disseminating digital data, regarding not only works of art and culture but also content relevant to these works, such as reportorial and academic articles, reviews, opinion pieces, event or exhibition retrospectives, and more. This study has shed a light on the level of integration of Semantic Web technologies and how the various metrics that quantify different Semantic Web technologies can be used not only to assess Semantic Web integration, but might also influence or predict website popularity to a small extent.

According to the findings, many of the distributions of the various variables displayed a pattern of having two peaks, one at the lowest and one at the highest value. This indicates that most websites either completely ignore the use of a specific Semantic Web technology or fully commit to it, implementing it comprehensively. Additionally, the moderate correlations between the various metrics indicated that integration with the Semantic Web as a general goal is mostly either ignored or pursued thoroughly. Finally, through the Gradient Boosting analysis it was established that the integration of schema.org structure data in a website was the most important factor in the ability to predict the website's popularity.

The research presented in this study was limited in its ability to fully include all relevant websites. Additional insight might be found in websites that might not have been discovered by the study's methodology or that were excluded for not being in English. Further research, monitoring Semantic Web integration in the websites of developing countries such as China or India might produce different results and assist in creating a more comprehensive overview of the landscape of Semantic Web technologies integration. Additionally, the present research focused exclusively in the areas of art and culture, but things might be different in other fields. The line of research presented here can continue in the future, with the focus shifting from media relating to art and culture to media relating to other fields such as sports, technology, consumer products, and more. The Semantic Web Technologies Integration rating introduced is content-agnostic and as such can be used to evaluate integration in any field. Additionally, its simplicity allows its use even without the automated crawling algorithm described in this article, as long as the data set of relevant websites is small. Enriching the data-gathering process with even more technologies that encompass aspects of the Semantic Web as they become popular in the future is also important and can form a basis for future research.

Studying and analyzing the tangible presence of the Semantic Web is an important step in evaluating its progress and can be of valuable help in achieving its true potential, which so far remains largely untapped. The increased relevance of social media and the marketing importance of SEO can both become incentives to further expand both the quantity and the quality of machine-readable structured rich data in websites of any magnitude or topic through technologies such as Open Graph and Schema.org. Furthermore, new challenges emerge with the decentralization principles brought forward with the popularization of blockchain technology and the Semantic Web must rise to meet them in order to expand and encompass all aspects of the World Wide Web as it evolves with unprecedented celerity.

Author Contributions: Conceptualization, A.G.; Formal analysis, M.P., N.K. and A.K.; Investigation, A.G., M.P. and A.L.; Methodology, A.G. and M.P.; Project administration, A.G.; Resources, N.K. and A.K.; Software, M.P.; Supervision, A.G. and I.V.; Visualization, A.L.; Writing—original draft, M.P., N.K. and A.K.; Writing—review and editing, M.P. and I.V. All authors have read and agreed to the published version of the manuscript.

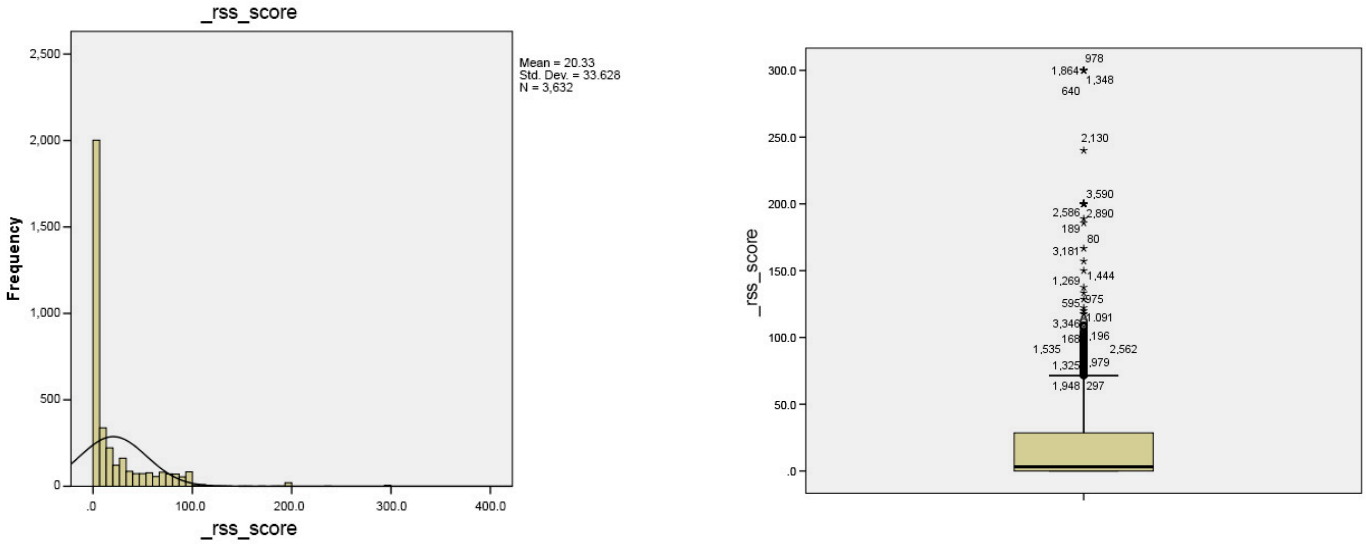
Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available in Zenodo at [10.5281/zenodo.5811988], reference number [10.5281/zenodo.5811988].

Conflicts of Interest: The authors declare no conflict of interest.

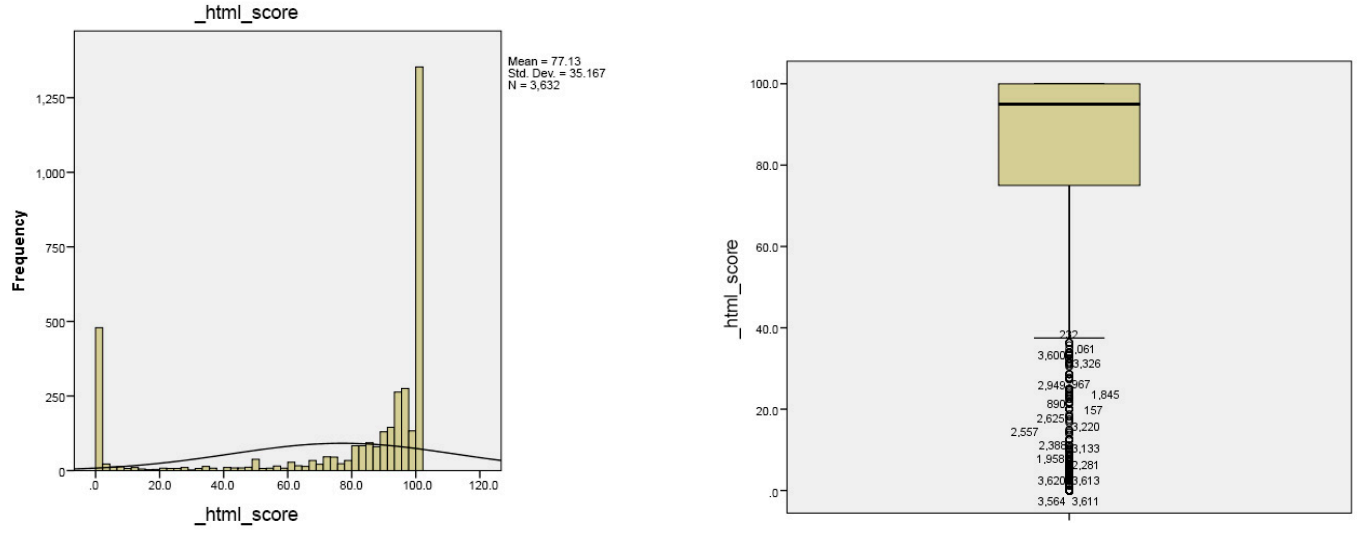
Appendix A

This appendix presents the histogram and boxplots for the variables _rss_score, _html_score, _og_score, _twitter_score, _schema_score, _html_variety, _og_variety, _twitter_variety and _swti. These distributions and dispersions are discussed in detail in Section 5 of this article.



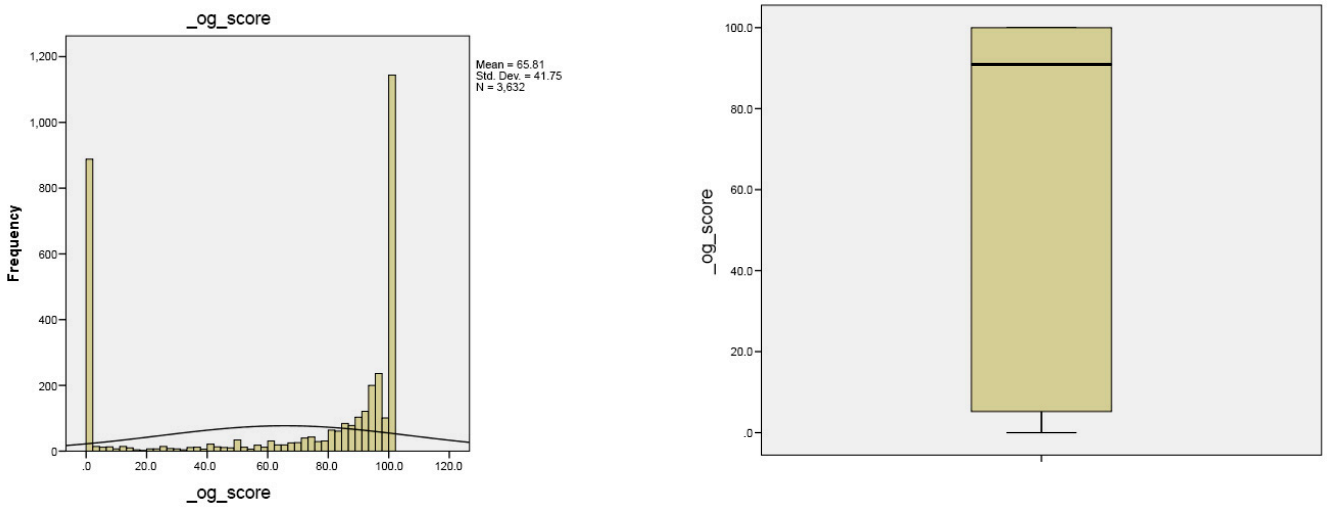
(**a**) (**b**)

Figure A1. The histogram (**a**) and boxplot (**b**) of the _rss_score variable showing a non-normal distribution.



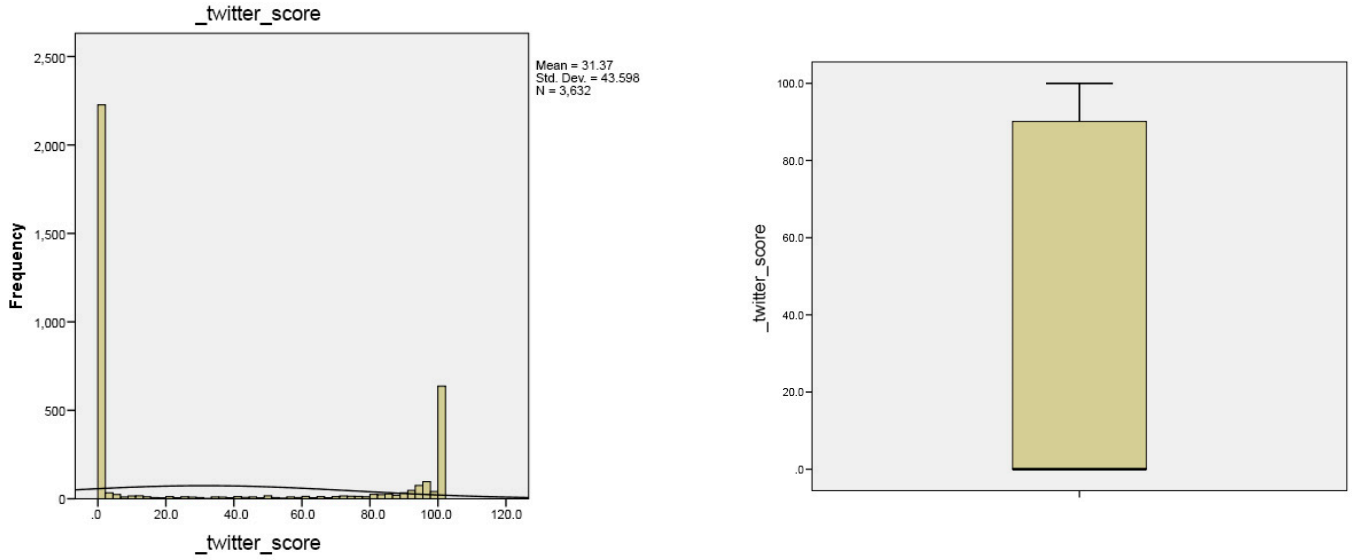
(**a**) (**b**)

Figure A2. The histogram (**a**) and boxplot (**b**) of the _html_score variable showing a non-normal distribution.



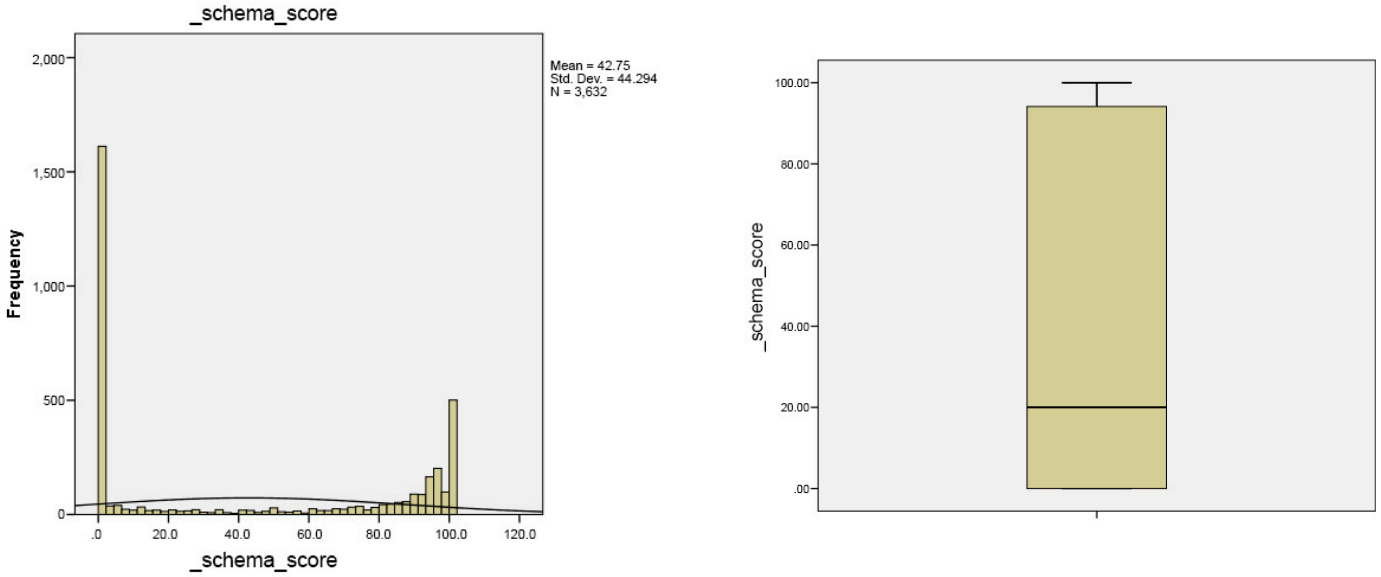
(**a**) (**b**)

Figure A3. The histogram (**a**) and boxplot (**b**) of the _og_score variable showing a non-normal distribution.



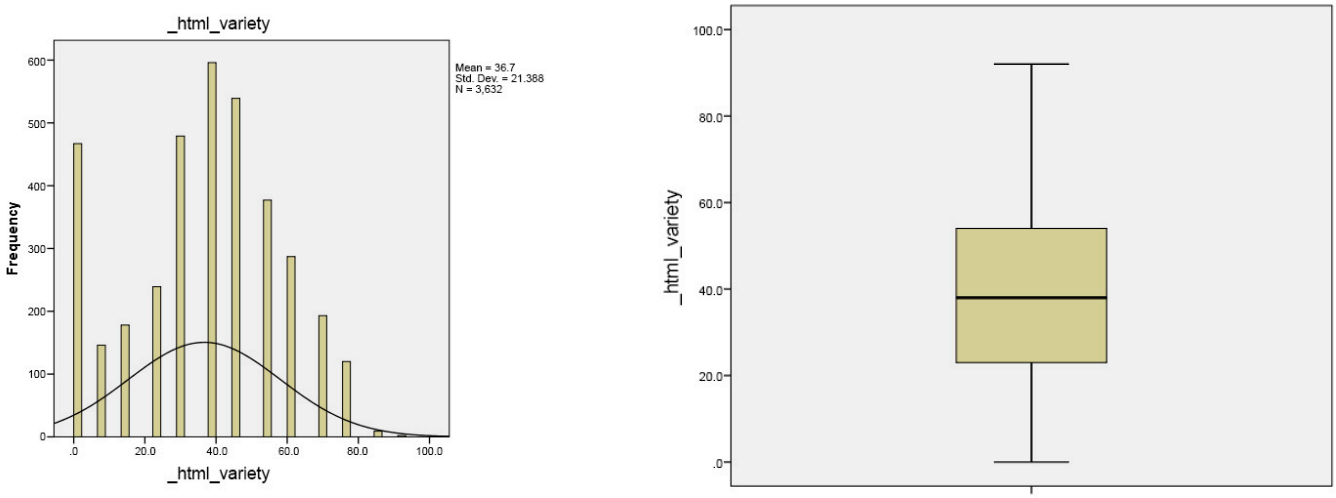
(**a**) (**b**)

Figure A4. The histogram (**a**) and boxplot (**b**) of the _twitter_score variable showing a non-normal distribution.



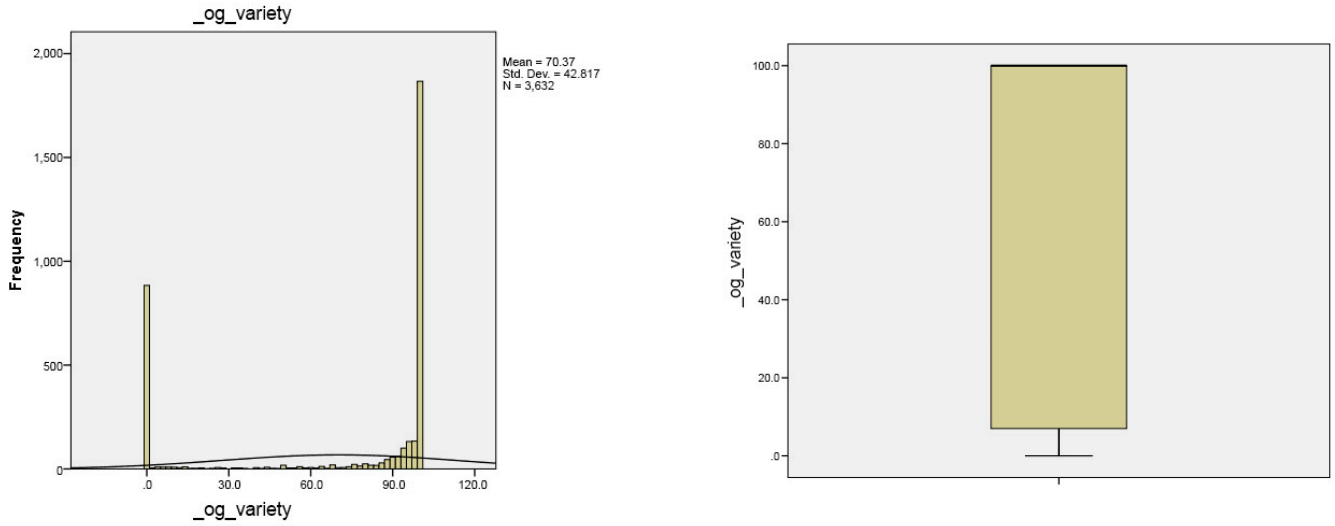
(**a**) (**b**)

Figure A5. The histogram (**a**) and boxplot (**b**) of the _schema_score variable showing a non-normal distribution.



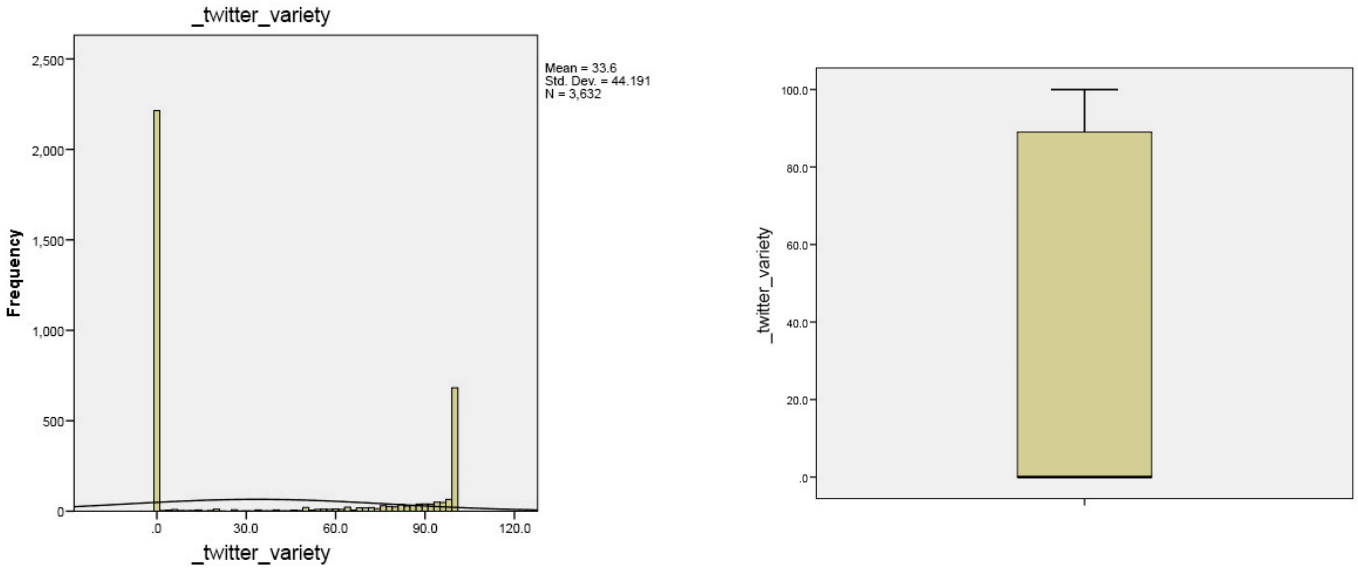
(**a**) (**b**)

Figure A6. The histogram (**a**) and boxplot (**b**) of the _html_variety variable showing a non-normal distribution.



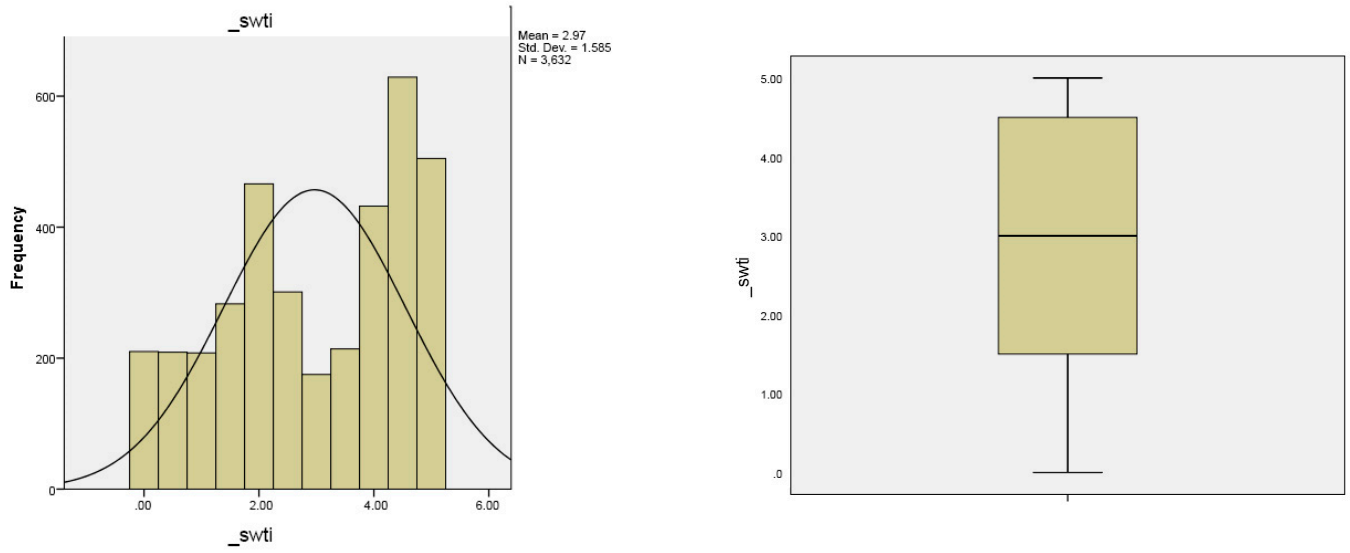
(**a**) (**b**)

Figure A7. The histogram (**a**) and boxplot (**b**) of the _og_variety variable showing a non-normal distribution.



(**a**) (**b**)

Figure A8. The histogram (**a**) and boxplot (**b**) of the _twitter_variety variable showing a non-normal distribution.



(**a**) (**b**)

Figure A9. The histogram (**a**) and boxplot (**b**) of the _swti variable showing a non-normal distribution.

References

1. Berners-Lee, T.; Hendler, J.; Lassila, O. The semantic web. *Sci. Am.* **2001**, *284*, 34–43. [CrossRef]
2. Bing, L.; Chan, K.C.; Carr, L. Using aligned ontology model to convert cultural heritage resources into semantic web. In Proceedings of the 2014 IEEE International Conference on Semantic Computing, Newport Beach, CA, USA, 16–18 June 2014; pp. 120–123.
3. Panagiotidis, K.; Veglis, A. Transitions in Journalism—Toward a Semantic-Oriented Technological Framework. *J. Media* **2020**, *1*, 1–17. [CrossRef]
4. Heravi, B.R.; McGinnis, J. Introducing social semantic journalism. *J. Media Innov.* **2015**, *2*, 131–140. [CrossRef]
5. Lim, Y.S. Semantic web and contextual information: Semantic network analysis of online journalistic texts. In *Recent Trends and Developments in Social Software*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 52–62.
6. Dertouzos, M. *The Unfinished Revolution: Human-Centered Computers and What They Can Do for Us*; HarperCollins: New York, NY, USA, 2001.
7. Nath, K.; Dhar, S.; Basishtha, S. Web 1.0 to Web 3.0—Evolution of the Web and its various challenges. In Proceedings of the ICROIT 2014—2014 International Conference on Reliability, Optimization and Information Technology, Faridabad, India, 6–8 February 2014; pp. 86–89. [CrossRef]
8. Varlamis, I.; Giannakoulopoulos, A.; Gouscos, D. Increased Content Accessibility For Wikis And Blogs. In Proceedings of the 4th Mediterranean Conference on Information Systems MCIS, Athen, Greece, 25–27 September 2009.
9. Belk, M.; Germanakos, P.; Tsianos, N.; Lekkas, Z.; Mourlas, C.; Samaras, G. Adapting Generic Web Structures with Semantic Web Technologies: A Cognitive Approach. In Proceedings of the 4th International Workshop on Personalised Access, Profile Management, and Context Awareness in Databases (PersDB 2010). 2010. Available online: http://www.vldb.org/archives/website/2010/proceedings/files/vldb_2010_workshop/PersDB_2010/resources/PersDB2010_6.pdf (accessed on 30 December 2021).
10. Blake, J. On defining the cultural heritage. *Int. Comp. Law Q.* **2000**, *49*, 61–85. [CrossRef]
11. Kabassi, K. Evaluating museum websites using a combination of decision-making theories. *J. Herit. Tour.* **2019**, *14*, 1–17. [CrossRef]
12. Kiourexidou, M.; Antonopoulos, N.; Kiourexidou, E.; Piagkou, M.; Kotsakis, R.; Natsis, K. Multimodal Technologies and Interaction Websites with Multimedia Content: A Heuristic Evaluation of the Medical/Anatomical Museums. *Multimodal Technol. Interact.* **2019**, *3*, 42. [CrossRef]
13. Dannélls, D.; Damova, M.; Enache, R.; Chechev, M. A framework for improved access to museum databases in the semantic web. In Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage, Hissar, Bulgaria, 16 September 2011; pp. 3–10.
14. Dimoulas, C.; Veglis, A.; Kalliris, G. Audiovisual Hypermedia in the Semantic Web. In *Encyclopedia of Information Science and Technology*, 3rd ed.; IGI Global: Hershey, PA, USA, 2015; pp. 7594–7604. [CrossRef]
15. Janssen, S. Art journalism and cultural change: The coverage of the arts in Dutch newspapers 1965–1990. *Poetics* **1999**, *26*, 329–348. [CrossRef]
16. Matthews, B. Semantic web technologies. *E-learning* **2005**, *6*, 8.
17. Kumar, S.; Chaudhary, N. A Novel Trust Scheme in Semantic Web. In *Information and Communication Technology for Sustainable Development*; Springer: Singapore, 2020; pp. 103–110.
18. World Wide Web Concortium (W3C). Available online: https://www.w3.org/ (accessed on 20 December 2021).

19. Ning, X.; Jin, H.; Wu, H. RSS: A framework enabling ranked search on the semantic web. *Inf. Process. Manag.* **2008**, *44*, 893–909. [CrossRef]
20. Necula, S.C.; Păvăloaia, V.D.; Strîmbei, C.; Dospinescu, O. Enhancement of e-commerce websites with semantic web technologies. *Sustainability* **2018**, *10*, 1955. [CrossRef]
21. Patel-Schneider, P.F. Analyzing schema. org. In *International Semantic Web Conference*; Springer: Cham, Switzerland, 2014; pp. 261–276.
22. Meusel, R.; Bizer, C.; Paulheim, H. A web-scale study of the adoption and evolution of the schema. org vocabulary over time. In Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics, Larnaca, Cyprus, 13–15 July 2015; pp. 1–11.
23. Mika, P. On schema. org and why it matters for the web. *IEEE Internet Comput.* **2015**, *19*, 52–55. [CrossRef]
24. Common Crawl. Available online: https://commoncrawl.org/ (accessed on 20 December 2021).
25. Giannakoulopoulos, A.; Pergantis, M.; Konstantinou, N.; Lamprogeorgos, A.; Limniati, L.; Varlamis, I. Exploring the Dominance of the English Language on the Websites of EU Countries. *Future Internet* **2020**, *12*, 76. [CrossRef]
26. Alonso, I.; Bea, E. A tentative model to measure city brands on the Internet. *Place Branding Public Dipl.* **2012**, *8*, 311–328. [CrossRef]
27. Govers, R.; Van Wijk, J.; Go, F. Website Analysis: Brand Africa. In *International Place Branding Yearbook 2010: Place Branding in the New Age of Innovation*; Palgrave Macmillan: London, UK, 2010; pp. 156–171. [CrossRef]
28. Alexa Internet-About Us. Available online: https://www.alexa.com/about (accessed on 20 December 2021).
29. PHP cURL Introduction. Available online: https://www.php.net/manual/en/intro.curl.php (accessed on 20 December 2021).
30. Powers, S. *Practical RDF*; O'Reilly Media, Inc.: Cambridge, MA, USA, 2003; p. 10.
31. HTML Semantic Elements. Available online: https://www.w3schools.com/html/html5_semantic_elements.asp (accessed on 20 December 2021).
32. The Open Graph Protocol. Available online: https://ogp.me/ (accessed on 20 December 2021).
33. About Twitter Cards. Available online: https://developer.twitter.com/en/docs/twitter-for-websites/cards/overview/abouts-cards (accessed on 20 December 2021).
34. Microformats—Building Blocks for Data-Rich Web Pages. Available online: https://microformats.org/ (accessed on 20 December 2021).
35. Roussos, P.L.; Tsaousis, G. *Statistics in Behavioural Sciences Using SPSS*; TOPOS: Athens, Greece, 2011.
36. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
37. Zhang, L.; Zhan, C. Machine learning in rock facies classification: An application of XGBoost. In Proceedings of the International Geophysical Conference, Qingdao, China, 17–20 April 2017; Society of Exploration Geophysicists and Chinese Petroleum Society: Tulsa, OK, USA, 2017; pp. 1371–1374.
38. Steinberg, D. Classification and regression trees. In *The Top Ten Algorithms in Data Mining*; Wu, X., Kumar, V., Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2009; pp. 179–202.
39. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
40. Caruana, R.; Niculescu-Mizil, A. An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 161–168.
41. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]