Evolution of Legal Statements on the Web

Holger M. Kienle
University of Victoria, Canada
hkienle@acm.org

Crina A. Vasiliu
University of Victoria MBA Alumni, Canada
crina.vasiliu@alumni.uvic.ca

Abstract

In this paper we propose to study the evolution of legal statements that can be found in web sites. Legal statements are an important part of each web site because they can be seen as a contract between the owner of the site and its users. For example, a site’s privacy policy explains what kind of data is collected from users by the operator and how it is processed. Operators use terms of use to put restrictions on the conduct of users. In this paper we describe our proposal for a research agenda and methodology that analyzes the evolution of legal statements on the web. The research agenda argues that studying the content of legal statements and how they change over time allows to analyze and understand the evolution of the web from different viewpoints. Specifically, changing legal statements allow to identify emerging legal developments, to expose shifting business objectives, and to track the balance of power between operators and users.

Our suggested methodology proposes to obtain historical snapshots of web sites available in the Internet Archive, to group them into different classes, and to analyze the content of the legal document as well as to compute metrics such as size and readability scores. The obtained data can then be used to formulate hypotheses about the evolution of certain characteristics of the web. We discuss a pilot study that instantiates our methodology. This study is based on five snapshots of 15 different web sites, and it shows that the methodology is feasible and can generate meaningful results.

Keywords: Internet, legal factors

1. Introduction

The goal of this research is to study the contents of both terms of use and privacy policy statements on web sites as well as their evolution. These kind of statements are typically available from hyperlinks displayed at the bottom of a web page. While such statements can now be found on many web sites, they have not been present during the early days of the web. The site of General Electric, for instance, had no privacy statement prior to June 1997, and no terms and conditions prior to April 1999.

The legal statements of web sites can be seen as a contract between the owner of a web site and its users. The terms of use address limitations on the use of materials on the site (e.g., downloading of copyrighted contents for personal, non-commercial purposes only), trademarks that are claimed by the site owner and other organizations, disclaimers of liability (e.g., no guarantee of the accuracy of information), and restrictions of linking to the site (e.g., only links to the home page are allowed, but no deep-linking). The privacy policy of a web site addresses issues surrounding the storage, processing, and transfer of both personal information such as birthday, social security number, email, answers to questionnaires, user preference settings, and (flash) cookies.

Nowadays virtually every major web site has legal statements and as a direct consequence large portions1 of the web—and therefore the web’s users—are governed by legal contracts. Thus, it is important to gain a better understanding of the meaning of these statements, how they impact the web and its users, and their historical development with the goal to project future trends. Our research can be seen as one approach to analyze and understand socio-technical issues surrounding the web.

Indeed, large portions of the legal statement can be seen as restrictions imposed by the web site operators on the visitors of their site. However, it is important to understand that legal statements reflect the wishes and desires of the content provider, but do not indicate which of these are actually grounded in law and can be enforced in front of a judge. The aim of this paper is not to provide an analysis of the legal justifiability of the statements made on web sites.

1.1. Organization of the Paper

Section 2 outlines our research agenda where we propose to process legal statements as one approach to analyze cer-

---

1The percentage of web pages under legal statements may not be large in terms of the WWW’s total page size, but certainly in terms of the number of clicked pages.
tain aspects of the evolution of the web. Section 3 identifies related work in the areas of web site analysis as well as electronic processing of legal content. To our knowledge, this is the first kind of work that addresses the evolution of the web from a legal perspective. In Section 4, we describe our envisioned methodology on how a study of the evolution of legal statements on the web can be conducted. We have implemented part of the methodology in a preliminary study. The goal of this study, presented in Section 5, was to gather first evidence that our research agenda is feasible and can yield interesting results. The paper closes with future work (Section 6) and conclusions (Section 7).

2. Research Agenda

In this section we outline why we propose to analyze the way that legal statements of web sites have evolved over time. We outline why we believe that it is worthwhile to pursue this research and we hope that it will encourage other researcher to continue following this path.

Specifically, we believe that legal statements can be used as input to track the evolution of certain characteristics of web sites from different viewpoints. Understanding the evolution of (classes of) web sites may also enable us, by extrapolation, to better understand the evolution of the WWW as a whole.

The following sections give examples of viewpoints that may be used when analyzing the evolution of legal statements.

2.1. Tracking of Legal Developments

There is the perception that the web is becoming increasingly regulated. For instance, Baker says in the foreword of Isenberg’s book: “No longer an information ‘wild, wild west,’ the Internet increasingly is influenced by legal considerations” [14, p. xi]. In this book, Isenberg says that U.S. courts did not have to deal with Internet-related issues until 1994, but in 1999 there were more than 550 published court opinions; in 2001 this number then more than doubled to over 1,100 opinions [14].

We believe that analyzing the content of legal statements and their evolution over the years can be used to track legal developments, because such statements probably change as a reaction to new laws and legal proceedings. For example, the Health Information and Portability Accountability Act (HIPAA) of 1996 mandated that the U.S. government had to come up with regulations for health information privacy. As a result, health care provides have to comply with certain rules starting April 2003; these rules also have to be reflected on the provider’s web site privacy policies [4].

Note that not all legal developments are necessarily reflected in the legal statements. For example, the implementation of accessibility guidelines (e.g., Section 508 of the U.S. Rehabilitation Act) only impacts how the web site is build. However, some organizations choose to make their commitment to accessibility explicit. For example, General Electric has a link to the GE Accessibility Statement at the bottom of each page.

2.2. Exposure of Business Policies and Objectives

Organizations may decide to change or clarify their business policies and objectives in light of new legal developments or business drivers. An example of a business driver that impacts legal statements is the concept of information asset, which is information of value (e.g., patient records, intellectual property, or customer information) that is owned and/or used by an organization and needs to be protected [28].

Besides business drivers, court cases and decisions can have an impact on legal statements. Such an example is the practice of deep-linking, which led to a widely publicized legal action in 1999. In TicketMaster vs. Tickets.com, TicketMaster sued a competitor, Tickets.com, for using deep links to their event pages [11]. A U.S. district court, as part of a motion to dismiss this case, came to the conclusion that deep-linking is not a violation of copyright. The practice of framing a site’s page by a different site (such as in Washington Post vs. TotalNews) is another example of a mechanism that organizations may want to control [27].

Regardless of the outcome of particular legal proceedings, which can take for years, they probably prompted other web site operators to rethink if and how they wanted to restrict deep-linking and framing of external sites. As a consequence, legal statements are modified to explicitly address operators’ business policies. In the first case, Tickets.com claimed it was losing revenue because visitors were going directly to the page they needed, and were skipping pages that contained advertising. In the second case, TotalNews made revenue by posting advertisements next to copyrighted content belonging to external sources.

2.3. Power Balance Among Stakeholders

For (traditional) web sites, we can distinguish between two stakeholders: content providers (i.e., the web site operator who makes information available) and content consumers (i.e., the users of the web site) [15]. Web sites that have characteristics of the so-called Web 2.0 [20] blur this distinction because content consumers can be also content providers.

The relationship between the web site operator and its users is governed on the one hand by the content and interactions that the web site offers (e.g., users may have the option to comment on information provided by the operator or other users), but also by the legal statements (e.g., offensive
comments and violations of copyright are not permissible). Thus, analyzing the content of legal statements gives an understanding of the rights and obligations—in other words, the power balance—between operators and users. Generally, this balance is biased towards operators because they can draft (and later on modify) the legal statements without consulting users. On the other hand, users have the freedom to abandon a web site.

There are examples of questionable practices by operators that have resulted in user protests or legal action. For example, in 2000 the failed online retailer Toysmart.com planned to sell personal customer information (supposedly in violation of its own privacy policy). Another example that sparked controversy over privacy is Facebook Beacon, which tracked purchases of users and made them visible to others.

Analysis of the evolution of legal statements is important also because as the complexity of web sites increases (e.g., because of Web 2.0 features), legal statements increase in complexity as well. Antón and Earp say that “each [privacy] policy is different, and it is difficult—if not impossible—for the average user to compare and comprehend these policies” [1]. In fact, web sites offering functionality that matches complexity and functionality of shrink-wrapped software (e.g., Google Gmail and Adobe Photoshop Express) have legal statements that incorporate elements of shrink-wrapped licenses (e.g., statements that prohibit to decompile or reverse engineer the software).

3. Related Work

Many researchers have studied web sites with the goal to characterize their properties and their structures. In early work, Warren et al. propose to adapt hypertext metrics for web sites [34]. They show results for four smaller web sites, using metrics such as number of links and tags, cyclical complexity, and mean fan-in/out. Linos et al. use their own web crawler, Perlbot, to collect information from 10 university web sites in different countries [16]. They report data about percentage of broken links and non-HTTP links, number of Javascript and cookies, and age distribution of files. Zhou et al. have analyzed 18 web sites for topological properties (degree distribution, nearest-neighbors, rich-club connectivity, etc.) to characterize graphs [35]. Besides purely numerical approaches, researchers have also investigated the structure of web sites. Martin and Martin use Rigi to visualize the (higher-level) link structure of web sites [17]. Hassan and Holt have extended the Portable BookShelf to visualize the architecture of web site. Ricca visualizes the structure of multilingual sites, identifying inconsistencies in link structures between languages [22].

The above work is primarily focused on technical characteristics of web sites—that is, results are primarily obtained by analyzing HTML and other code artifacts. In contrast, our work is concerned with the web site’s content. Specifically, we propose metrics related to the site’s legal statements in terms of length (number of words and sentences) and readability scores. The structure of the site is of no or little concern for our research since the legal statements that we target are typically contained in a few self-contained HTML pages. Other researchers have also focused on the content of web sites. For example, Tonella et al. extract natural-language keywords from web pages and then use them to generate a clustering by grouping together pages that have common keywords [30]. The same authors use simple natural language processing (NLP) techniques to analyze the contents and alignment of multilingual sites for restructuring [29]. First, the language of each web page is determined based on matching words in dictionaries. Then, pages in different languages are aligned based on a score that assumes that each word in one page may have a translation in the other. The alignment makes use of stop words such as prepositions and articles, and applies morphological transformations to match words.

The approaches described so far have in common the fact that they operate on a single snapshot of a web site. In contrast to this, researchers have also studied multiple snapshots of the same site. This is typically done with the goal to show historical trends. For example, WebCiao is an environment to analyze and visualize the structure of web sites [7]. It offers a differencing tool that can compare two snapshots of a site. The result can then be visualized as a color-coded graph that shows which pages have been changed, deleted, and unmodified with respect to the old snapshot. The ReWeb tool provides a graph-based visualization of the link structure of multiple snapshots of a web site [23] [24]. Different snapshots are assigned different colors. There is also a visualization that summarizes the amount of changes for each snapshot, allowing to spot major restructurings easily. Bailey and Burd have conducted a study about the evolution of web sites in the UK with respect to accessibility [5]. They leverage Access Varlet’s accessibility report to compute an accessibility score for the web sites over a period of five years (one snapshot as early in the year as possible taken from the Internet Archive). The results are analyzed by grouping the sites into higher education, public organizations, and commercial companies. Similarly, Fukuda et al. show data on accessibility for blind users of seven major web sites (randomly selected) over 8 years (1997–2004) [10]. Toyoda and Kitsuregawa study the evolution of web communities (i.e., web pages that address a certain topic that is of common interest to a number of people) [31]. They used their own crawler to collect four repositories of collections of web pages (from 2000 to 2002, ranging from
17 to 45 millions of pages). Based on this data, they analyze how communities emerge, split, merge and dissolve. They also discuss size and growth rate of communities.

To our knowledge, there is no related work that analyzes a series of snapshots of legal statements on web sites. However, there is related work that analyzes legal texts. An interesting related field is legal informatics, which encompasses the (natural-language) processing of legal documents. For example, Antón and Earp use goal mining to (manually) identify goals of privacy policies on the web. They have analyzed 23 web sites of the healthcare domain and identified 131 unique goals, which are expressed in structured natural language [3]. They have also implemented a web-based goal management tool that contains over 1,200 extracted goals from over 100 legal statements on the web [6]. The EULA Analyzer is an online tool that analyzes End-User License Agreements (EULAs) with the goal to identify clauses that are of particular concern to users that have to agree to these licenses. Once the agreement is pasted into a text box, the analyzer provides metrics such as word count, number of sentences, and readability scores. Furthermore, selected sentences are highlighted and annotated to provide guidance for humans. A similar tool is the Java-based EULAlizer. There is also work on analyzing software licenses. Tuunanen et al. present a tool that analyzes the build process of an existing software system, identifying (open-source) components along with their licenses that are part of the software [32]. Licenses are identified based on text searching with regular expressions.

4. Study Methodology

In this section we outline how we plan to conduct our study (i.e., methodology, retrieval of historical snapshots, data processing and metrics calculations, and selection and classification of web sites).

4.1. Grounded Theory

We propose using a grounded theory [12] approach to generate and analyze results. In grounded theory, an existing, predetermined subject (in our case, legal statements of web sites) is analyzed to gain a better understanding of its underlying characteristics. This is achieved by the systematic gathering, analyzing, and classifying of data. The data is subsequently used to develop theories or form hypotheses. Thus, results are not obtained based upon preconceived ideas, but are generated bottom-up from the available data. We believe that this approach is most suitable for new research that has not yet developed a foundation of accepted results or theories.

4.2. Internet Archive as Data Source

To analyze the evolution of the web sites we are using the Internet Archive’s Wayback Machine, which archives snapshots of web sites starting from the year 1996. This approach has been used by other researchers before to obtain historical data (e.g., [5] [10] [13]).

Unfortunately, the amount and sequence of snapshots varies considerably depending on the web site and year. Generally, earlier years have fewer snapshots than later years. Accessing the Wayback Machine poses some difficulties because of unreliable connections, inaccessible or partial versions of sites, delays of page downloads because of missing Java applets (General Electric site), etc. Furthermore, an operator may block the crawling of their site via a robots.txt file (e.g., exxonmobile.com), which the archive respects.

4.3. Processing of the Legal Statements

Our extraction of the legal statements from a selected snapshot is currently done mostly manually. We browse the home page of the site and look for relevant links, which are typically located at the bottom of the page. Once we identified the URLs that contain the legal statements, we save them in a textual file (one line per URL) to enable subsequent automatic processing.

The processing of the legal statements is done as follows. We first download them with wget, clean up the HTML with tidy (more relevant for older pages), and then transform the HTML to plain ASCII with the html2text tool. Subsequent analyses and metrics calculations are based on the ASCII versions.

4.4. Selection and Classification of Web Sites

Generally, the number and distribution of web sites available on the Internet cannot be determined and thus selecting a set of sample web sites is bound to be biased. For our current study we do not aim for a sample that represents the Web as a whole, but rather try to select a suitable number of sites that allow us to expose and contrast historical trends for certain domains.

For selecting web sites we used the following restrictions. First, all sites are based in the U.S. and primarily target the population of this country. This limitation allows us to focus on one particular legal system and its relevant legal cases. Second, we decided to investigate larger web sites that have a significant number of users. We assume that larger sites have a particular interest in developing strategies to deal with legal and policy issues. Thus, it can be expected that their legal statements are frequently re-evaluated.

---

4http://www.spywareguide.com/analyze/
5http://www.javacoolsoftware.com/eulalyzer.html
6http://www.archive.org/web/web.php
in light of new legal requirements and consumer demands. Also, larger web sites are more likely to have a sufficient number of historical data for our purposes. They also reduce the risk of tainted domains [18] where the owner of the domain has changed in the meanwhile. Third, in order to have sufficient historical data we only selected sites that were archived from 1998–2007 with at least one snapshot per year. As a result, our study does not cover recent popular Web 2.0 sites such as YouTube and Wikipedia.

For our study, we group Web sites into the following groups:

**brick-and-mortar businesses:** These kind of businesses are traditional companies that put information on the Internet, but do not engage in business to consumer (B2C) activities [21] [26]. Typically, consumers can get information about the company (e.g., press releases and stock information), its products, and how to obtain them. We considered candidates based on Fortune Magazine’s Fortune 1000 list of 2008, which provides a ranking of the largest corporations in the U.S.

**e-businesses:** These businesses are retailers that sell products online. They may be pure online retailers (e-tailers) or have a clicks-and-bricks hybrid business model [21] [26]. We considered businesses that are listed in Alexa’s U.S. Top 100, and the worldwide Top 500 (April 2008). Alexa (alexa.com) collects information about visited sites from users that have installed the Alexa Toolbar on IE or Firefox. The ranking should be taken with a grain of salt because of sampling bias.

**universities:** We considered universities in the U.S. that are members of the Association of American Universities (AAU). As of 2007 there are 62 members. This choice introduces a bias towards more established and prestigious universities.

Note that we do not claim that the above groups are representative of the web as a whole. For example, we are (currently) not looking at the following classes: web portals, social networks, online media, government, and pornography. Also, the above groups represent major entities whose characteristics do not necessarily correspond to smaller entities (e.g., large companies vs. SMEs).

### 4.5. Analysis of Snapshots

For analysis we propose to look at both metrics-based and content-based approaches. To track the evolution of the legal statements we can analyze their

**change frequency:** The change frequency of legal statements can give an indication of their dynamism. Generally, the web is highly dynamic in terms of URLs that disappear and change over time [31], but often the contents of a page change only in their markup or in trivial ways [9]. It would be interesting to see whether this holds for legal statements as well.

**structure-based metrics:** Metrics can be used to assess and track certain characteristics of legal statements. An example of a simple metric is the size of the legal document in terms of words (cf. Section 5.1). Looking at the historical development of the size of a web page may be significant because generally “past changes to a page are a good predictor of future changes” [9]. Examples of a more sophisticated type of metric are readability scores (cf. Section 5.2).

**legal topics:** The above approaches do not take into account the meaning of the contents. A simple approach for analyzing legal contents is to manually skim through the texts to determine which legal topics are covered (cf. Section 5.3). Examples of topics are licensing, warranty and disclaimers, link policies, copyright and trademark, privacy, applicable jurisdiction, etc. The legal aspects are not pre-determined in advance, but are extended and re-organized (e.g., in a taxonomy or ontology) based on preliminary findings while analyzing the sites. This approach is in line with the spirit of grounded theory.

### 5. Pilot Study

In this section we report first results that we obtained with a smaller case study. The purpose of this study was to validate that our proposed methodology is feasible and that the analyses can provide useful results.

In this study, we have selected 15 web sites with snapshots from five years in two year increments (1998, 2000, 2002, 2004, and 2006). Because of the delay that the Internet Archive publishes its snapshots, 2008 is not available yet. The following web sites have been selected (five in each group, cf. Section 4.4):

**traditional businesses:** General Electric (ge.com), General Motors (gm.com), United Healthcare (uhc.com), Tyson Foods (tyson.com), and Valero Energy (valero.com). These businesses are listed in the Forbes list at positions 6, 4, 25, 88 and 16, respectively.

**e-businesses:** Adobe (adobe.com), Barnes & Noble (bn.com), Dell (dell.com), JP Morgan Chase (chase.com), and Kmart (kmart.com). These
businesses are listed in the Forbes list at positions 651, 446, 38, 12 and 45 (Sears Holdings), respectively. Some of them are also listed in the U.S. Top 100: Adobe (57), Dell (64), and JP Morgan Chase (44). In some cases we had to settle for a less popular site because we could not obtain all snapshots that we needed from the Internet Archive. We selected Barnes & Noble instead of Amazon, and Kmart instead of Target, BestBuy, or Wal-Mart.

universities: Harvard University (harvard.edu), New York University (nyu.edu), Stanford University (stanford.edu), University of Buffalo (buffalo.edu), and University of Texas at Austin (utexas.edu). We noticed that many of the web sites of the AAU universities do not have legal statements. In order to obtain historical data for this group, we made the decision to consider only such sites as candidates that currently (as of June 2008) have legal statements. As a result, this sample is strongly biased towards sites that emphasize legal considerations.

5.1. Evolution in Size

The evolution of the length of legal statements in terms of bytes, words, and sentences can indicate whether they tend to cover more legal topics and/or describe a legal requirement in more detail. Generally, longer legal statements indicate increasing complexity and as a result users may have more difficulty in comprehending them and are less likely to read them.

Figure 1. Evolution of the size of legal statements in words

Figure 1 gives a chart of the size evolution of the legal documents. The x-axis shows the five snapshots with increasing years to the right (1998, 2000, 2002, 2004 and 2006). The y-axis shows the accumulated size of the legal statements in words. The bars depict the size of all sites in each group: the universities (yellow), brick-and-mortar businesses (blue) and e-businesses (red). The green line gives the total count of all sites. The word count has been computed with `wc -w`.

The individual groups show a clear difference in size. University sites clearly have the least legal text. (Note that a random sample of university sites would show an even smaller size.) This leads to

hypothesis 1: University web sites tend to have comparatively little legal text, followed by brick-and-mortar companies. Sites with a lot of legal text tend to be e-businesses.

This makes sense intuitively because university sites are non-commercial in nature and target a comparably smaller group of users. The brick-and-mortar companies we looked at are large companies that have a high visibility and hence a strong interest in protecting and controlling their assets. Lastly, e-businesses have the same characteristics as brick-and-mortar companies, but their web sites have to support sophisticated user-interactions in order to support their business activities with consumers. In future work, this hypothesis could be extended to other groups. It would be also interesting to contrast different countries (e.g., Canadian vs. U.S. AAU member universities).

Figure 2. Regression curves of data in Figure 1

Looking at Figure 1, it can be observed that all groups are increasing in size over the years with a tendency of slowing growth. Figure 2 depicts the same data points as Figure 1 but superimposes regression curves over the data series. Table 1 summarizes the kinds of regression curves that we fitted (linear, logarithmic, and power) along with a measure of good fit, $r^2$, which ranges from 0 to 1, where values closer
<table>
<thead>
<tr>
<th></th>
<th>logarithmic</th>
<th>linear</th>
<th>power</th>
</tr>
</thead>
<tbody>
<tr>
<td>universities</td>
<td>0.9471</td>
<td>0.9432</td>
<td>0.9067</td>
</tr>
<tr>
<td>brick and mortar</td>
<td>0.9795</td>
<td>0.9827</td>
<td>0.9951</td>
</tr>
<tr>
<td>e-businesses</td>
<td>0.9951</td>
<td>0.9579</td>
<td>0.9788</td>
</tr>
<tr>
<td>all</td>
<td>0.9868</td>
<td>0.9677</td>
<td>0.9842</td>
</tr>
</tbody>
</table>

Table 1. Fit of different regression curves

To 1 indicate a better fit. For universities and e-businesses, the best fitting curves are logarithmic, whereas for brick-and-mortars it is a power curve. The data set for all sites is best fitted with a logarithmic curve, leading to

**hypothesis 2:** The evolution of the size of legal statements on the web follows a logarithmic trend. Early years show fast grow in size than later years.

In future work, this hypothesis has to examined with a much larger data sample to gain more confidence.

5.2. Evolution of Readability Scores

There are several well-known readability tests that determine how easy it is to read and comprehend a text. The SMOG formula assesses the educational level needed to understand a text [19]. It is computed with \( \sqrt{\frac{p}{s}} + 3 \), where \( p \) denotes the number of poly-syllables (i.e., three or more syllables) and \( s \) denotes the number of sentences. To compute SMOG values we use the GNU style tool, Version 1.11. We determine the SMOG for a given snapshot by concatenating all legal statements (i.e., typically the terms of use and the privacy policy) of the snapshot and computing a single SMOG value for it.

Figure 3 depicts a chart that shows the average SMOG values for each group of web sites over time: universities (yellow inverted triangles), brick-and-mortar businesses (blue boxes) and e-businesses (red diamonds). The total average for all groups is given as green triangles. These averages are in an interval between 13.51 and 14.46. According to the SMOG Calculator\(^9\) this interval corresponds to the New York Times and requires a college education level (SMOG 13-15). The first year has the worst readability score and does not quite fit with the other years. This can be partially explained by huge sentences that enumerate trademarks. For example, GM’s copyright and trademark information in 1998 has a sentence of 146 words that enumerates all of the trademarks that GM claims, resulting in a SMOG of 17.1.

Another popular readability measure is the Flesch Reading Ease Score (FRES). With FRES, lower numbers mean increasing difficulty. It is computed as \( 206.835 - 15.597 \frac{y}{w} - 1.015 \frac{s}{w} \), where \( y \), \( w \), and \( s \) denote the total number of syllables, words and sentences, respectively. According to Wikipedia, “Time magazine scores about 52, and the Harvard Law Review has a general readability score in the low 30s.”\(^10\) To compute FRES we use GNU style as well.

Figure 4. Evolution of FRES readability

Analogous to Figure 3, Figure 4 shows the evolution of the FRES values. The total averages (green triangles) for both charts show a rough conformance in the sense that for both averages the year 1998 has the worst readability, the years 2000, 2002 and 2006 have roughly the same readability, and 2004 has the second-worst readability. However, in contrast to size there is no trend of increasing or decreasing readability over the years. Antón et al. have also used FRES to analyze legal statements of healthcare-related sites [4]. They analyzed three sites of brick-and-mortar businesses (pharmaceutical companies) and six e-businesses (health insurances and online drugstores), reporting average FRES values of 38.3 (for Summer 2000) and 35.8 (for September 2003). While they report an increase in the complexity

\(^9\)http://www.wordscount.info/hw/smog.jsp
\(^10\)http://en.wikipedia.org/wiki/Flesch-Kincaid_Readability_Test
of legal statements based on these data points, it seems difficult to judge whether this is (statically) significant. In another similar study, Antón et al. reported an average FRES of 35.0 and 40.1 for three banks and four insurance companies, respectively [2].

About the legal statements that we and Antón et al. analyzed, it can be said that generally the obtained readability scores indicate that these are sophisticated texts. This leads to

**hypothesis 3:** Legal statements on the web are difficult to read and comprehend, requiring an education that is at least at college level.

The advantage of readability scores such as SMOG and FRES are that they can be automatically computed. However, they cannot assess the difficulty of the subject area of the text for a reader. One approach to obtain more meaningful results would be to conduct user studies that evaluate comprehension of the contents of legal statements [33] [8].

Absolute readability values should be taken with a grain of salt. Since tools have different algorithms to determine syllables, words and sentences, the results can differ significantly. To better understand and quantify the differences, we have used two different tools—GNU style and the SMOG Calculator—to compute SMOG values for the terms of use statements of the e-business group. Figure 5 shows the results. The SMOG Calculator has a significantly higher score (average of 16.9) compared to GNU style (average of 14.2). These differences are significant in the sense that the higher average corresponds to a university degree (SMOG 16, Atlantic Monthly) or even post graduate studies (SMOG 17–18, Harvard Business Review), and the lower average corresponds to a college degree (SMOG 13–15, New York Times). These observations lead to

**hypothesis 4:** Readability scores can give a rough indication of the complexity of a legal document, but they are not suitable to determine the required level of education because these results are highly tool dependent.

In future research, we plan to investigate the discrepancies between readability scores among tools in more detail and try to find a “canonical” formatting of the legal documents that minimizes the discrepancies.

### 5.3. Evolution of Terms of Use Statements

We have conducted a simple content analysis to better understand the evolution of selected aspects of the terms of use (ToU) statements. These statements contain legal issues that are (supposedly) binding to users that browse the web site or use advanced functionality of the site. We read through all ToU statements looking for occurrences of warranties/disclaimers as well as link policies. We did only check if this particular topic is addressed without going into a deeper legal analysis.

Figure 6 shows the evolution of ToU statements of the different groups (bars) and the total average (green line). As it can be seen, not all web sites have ToU statements for the snapshots we looked at. Kmart is the only example of a site that first introduced a ToU (in 1998) but then in the next snapshot abandoned it. (This can be explained with a major repositioning of the site around 2000.) It appears that ToU statements are “sticky”: once they got introduced, they stay (but mutate). We noticed some interesting differences between the groups. Most strikingly, only one university has a ToU statement (Stanford, first in 2000), whereas all sites of the other two groups had one by 2004. In our sample e-businesses have adopted ToU statements in a slightly more aggressive manner than bricks-and-mortars. These observations lead to

**hypothesis 5:** Bricks-and-mortars and e-businesses have been early adopters of terms of use statements and they are rapidly approaching full adoption. In contrast, universities have started out later in adopting them (they are laggards) and their rate of adoption is rather slow.

---

11For instance, depending on the tool a heading may be a separate sentence or folded into the sentence following the heading. The first case will probably result in an improved readability, the latter case will worsen the readability.
In future work, we plan to assess the rate of adoption of legal statements with Roger's theory of diffusion of innovations [25]. It remains to be seen if the Internet Archive reaches back in time far enough to show a complete S-curve, which is characteristic of an innovation's adoptions rate.

Figure 7. Percentage of web sites with disclaimers in their terms of use statement

Figure 7 shows the percentage of web sites with disclaimers/warranties in their ToU statements. Disclaimers typically try to waive or limit the site operator's liabilities. E-businesses (red bars) show a steady increase of disclaimers with full adoption by 2002. In contrast, not all brick-and-mortar businesses (blue bars) have adopted them. Notably, United Healthcare had a disclaimer in 2000 and 2002 but then abandoned it. Stanford is the only university that has a disclaimer (since 2000).

Figure 8. Percentage of web sites that address linking in their terms of use statement

Figure 8 shows the percentage of web sites that address linking in their ToU statements. Sites can have statements that address links that point to it (incoming links), and links within the site that point to other ones (outgoing links to third parties). ToU statements require, for instance, to contact the operator before linking to the operator's site and to use a certain name for the link (e.g., General Electric, starting in 2004). For outgoing links, ToU statements typically state that they are provided as a convenience and that following them is "at your own risk" (e.g., General Electric, starting in 2000). The chart shows that the average number of ToU statements that address linking are increasing steadily. It also appears that brick-and-mortar businesses (blue bars) have a higher tendency to address linking compared to e-businesses (red bars) and universities (yellow bars). There are a few examples of companies that have included outgoing linking but then dropped it afterwards: Tyson Foods no longer addressed it in 2006, Adobe addressed it in 2000 only, and Kmart did not address it in 2000/2002. Because of these cases there are declining averages for bricks-and-mortars from 2004 to 2006, and for e-businesses from 2000 to 2002.

Both legal topics—disclaimers and linking—show a tendency to increase over the years, but this trend is also broken. The trend is most clearly pronounced for disclaimers for e-businesses (cf. red bars in Figure 7) and the total averages for linking (cf. green line in cf. Figure 8). This leads to

**hypothesis 6:** Terms of use statements tend to increasingly address legal topics such as disclaimers and linking.

Similar to hypothesis 2, larger data samples are required to confirm or refute this hypothesis. Furthermore, it remains to be seen if this trend generalize indeed over all legal topics.

6. Future Work

For future work we want to concentrate on extending the pilot study into the following directions. First, we plan to take snapshots of sites once per year (1997–2008 in the best case) to have more fine-grained evolution data. Second, we want to increase the number of web sites for each of the three groups in order to have a more diverse sample. Third, we also would like to add more groups (cf. Section 4.4).

Furthermore, to make the handling of larger data samples practical, we plan to work on approaches that increase the accuracy, formality, and automation of analyses. For example, a more formal analysis would transform the legal texts into a structured representation such as a goal model, taxonomy, or ontology. Such a representation would then enable us to raise the level of automation in subsequent analyses. Other (semi-)automated approaches from NLP may be applicable as well such as term extraction. For example, a repeated occurrence of the term "link" (and its morphological variations) in a paragraph probably addresses policies for outgoing and/or incoming links of the web site. The precision of automated approaches could then be measured against a "gold standard" obtained from the smaller sample of manually-derived results from this study.

7. Conclusions

In this paper we have described a research proposal for analyzing the evolution of legal statements of web sites. Legal statements can be analyzed in several dimensions: change frequency, structure-based metrics, and content. We
propose to pursue a grounded theory approach that first generates data for these dimension and then uses the data in formulating hypotheses about the evolution of legal statements on the web.

We have conducted a first pilot study that applies our approach. The study is based on 15 web sites whose legal statements are tracked with five snapshots. We analyzed the statements with two kinds of metrics (size in words, and readability scores) and a content analysis of selected legal issues (linking and disclaimers/warranties). Based on obtained data we have formulated a number of hypotheses about the evolution of legal statements on the web. Our pilot study confirmed that our research approach is feasible and can facilitate the generation of hypotheses. Now, future work based on a more comprehensive data sample is needed to further refine and/or refute the stated hypotheses.

References


This paper has been published before under the following Creative Commons license: Attribution-Noncommercial-Share Alike 3.0 United States. The license is available here: http://creativecommons.org/licenses/by-nc-sa/3.0/us/.