Using the Wayback Machine to Mine Websites in the Social Sciences: A Methodological Resource

Sanjay K. Arora and Yin Li
School of Public Policy, Georgia Institute of Technology, Atlanta, GA 30332-0345. E-mail: {sanjayk.arora, yli1508}@gatech.edu

Jan Youtie
Enterprise Innovation Institute, Georgia Institute of Technology, Atlanta, GA 30308. E-mail: jan.youtie@innovate.gatech.edu

Philip Shapira
Manchester Institute of Innovation Research, Manchester Business School, University of Manchester, Manchester, M13 9PL, UK. E-mail: pshapira@mbs.ac.uk

Websites offer an unobtrusive data source for developing and analyzing information about various types of social science phenomena. In this paper, we provide a methodological resource for social scientists looking to expand their toolkit using unstructured web-based text, and in particular, with the Wayback Machine, to access historical website data. After providing a literature review of existing research that uses the Wayback Machine, we put forward a step-by-step description of how the analyst can design a research project using archived websites. We draw on the example of a project that analyzes indicators of innovation activities and strategies in 300 U.S. small- and medium-sized enterprises in green goods industries. We present six steps to access historical Wayback website data: (a) sampling, (b) organizing and defining the boundaries of the web crawl, (c) crawling, (d) website variable operationalization, (e) integration with other data sources, and (f) analysis. Although our examples draw on specific types of firms in green goods industries, the method can be generalized to other areas of research. In discussing the limitations and benefits of using the Wayback Machine, we note that both machine and human effort are essential to developing a high-quality data set from archived web information.

Introduction

The Wayback Machine, operated by the nonprofit Internet Archive and available at archive.org, offers the ability to retrieve historical website content. The Wayback Machine was launched in 2001 with 10 billion archived pages, following an earlier 5-year preliminary data collection effort. By December 2014, the Wayback Machine reported that it had archived 435 billion web pages worldwide. The earliest pages date back to 1996, with expansion in the scale and scope of its collections since the late 1990s through to the present with the global growth of websites and posts.¹

For social scientists, the Wayback Machine offers a valuable large-scale data source to analyze web information over time. There are many reasons why social scientists would wish to retrieve legacy web content: in our case, we are interested to explore how this web archive can be used in studying firm behavior, for example, to assess whether information about earlier firm strategies and practices can be associated with sales or job growth. Historical and current website information can serve as a complement to other sources of data. Some archived web information may substitute for items in a questionnaire: It is widely known that although questionnaires are a long-established means to gather information about current characteristics and behaviors of companies, response rates to survey requests are dwindling (Baruch, 1999). Archived websites also provide a

¹Number of web pages saved reported on http://archive.org/web/ (accessed December 18, 2014). Further information on the history and operation of the Internet Archive and Wayback Machine available at https://archive.org/about/.

Received July 28, 2014; revised December 18, 2014; accepted December 19, 2014

© 2015 The Authors. Journal of the Association for Information Science and Technology published by Wiley Periodicals, Inc. on behalf of ASIS&T • Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.23503

JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY, 67(8):1904–1915, 2016
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
complement to other commonly used unobtrusive data sources—for example, corporate databases, scientific publications, and patents—which contain historic data but are often limited with respect to what the data set covers, be it financial information in the case of corporate databases, or limited information about scholarly research in publications and applications in patents. Moreover, in some fields, such as our current research into U.S.-based small- and medium-sized enterprises (SMEs) in the green goods industry, only 20% have a publication or patent, with the vast majority having their know-how and trade secrets engineered into the product (Arora, Youtie, Shapira, & Li, 2014; Li, Arora, Youtie, & Shapira, 2014). Archived websites thus offer both a breadth of information about company practices and have the capability to present this information from prior periods.

However, using the Wayback Machine is not necessarily straightforward. Various methodological challenges are raised in accessing, analyzing and interpreting archived websites. These include standard social science issues such as field definition and bounding, sampling, and measurement, as well as specific challenges related to the Wayback Machine archive itself. This paper’s contribution to the early but growing literature on the Wayback Machine concerns the approaches we used to identify and overcome these methodological issues. In particular, we put forth a six-step process for accessing archived website information about SMEs, consisting of: (a) sampling, (b) organizing and defining the boundaries of the web crawl, (c) crawling, (d) website variable operationalization, (e) integration with other data sources, and (f) analysis. We conclude by pointing out the importance of automation and manual verification to maximize the value of the Wayback Machine for social science analysis.

The paper begins with a synoptic bibliometric review of the use of the Wayback Machine. This review is based on an analysis of publications indexed in Google Scholar that reference the Wayback Machine. These results show growing and broad-based use of the Wayback Machine over the past decade in computer science, information retrieval, and library and archival fields, but rather less use in social science analysis. These approaches are addressed in the subsequent sections of the paper. We present a process for making use of Wayback Machine archived websites for social science analysis. This process is applied to a data set of U.S.-based SMEs in green goods industries to exemplify its operation. The paper closes by acknowledging limitations in using archived websites along with a discussion of how we have addressed these limitations and what pathways for future research could be pursued to improve the use of archived websites.

Wayback Machine in Google Scholar

This section of the paper reviews available literature on the use of archived websites stored in the Wayback Machine. We used Google Scholar to identify research documents about the Wayback Machine. A search for “Wayback Machine” produced 4,898 documents listed in the period 2000 to 2013. We removed 238 of the entries that were just one-line citations and 2,158 that referred to particular U.S. Patent and Trade Office patent documents. That left 2,593 articles, books, and other papers about the Wayback Machine. An analysis of annual counts of these documents over time shows that following the official launch of the archive in 2001, Google Scholar documents that reference the Wayback Machine grew substantially through to 2003 (Figure 1). After a slight decline in 2004, there was steady growth in documents referring to the Wayback Machine. From 2004 to 2013, the annual number of books, articles, and other papers about the Wayback Machine rose at an

FIG. 1. Growth in Google Scholar articles and books about the Wayback Machine. Annual counts of articles, books, and other documents. Search of Google Scholar performed May 2014. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]
Thirty-seven percent of these documents had sufficient information in the journal name to enable classification into fields. Thirty-one percent of the journal names fall into the information technology area, 16% into archive or library areas, and 11% into legal areas. The most highly cited information technology work is Resnik and Smith’s (2003) “The Web as a Parallel Corpus,” which has attracted more than 400 citations in Google Scholar. This work describes an effort to test web mining software using information from the Wayback Machine. Another highly cited work in information technology, which attracted more than 140 citations, is Baker et al. (2006) about the demands that the Wayback Machine and other web-based applications place on long-term storage and approaches to address these demands. A third highly cited work concerning information technology is “Riding the Waves of Web 2.0,” which is a summary of the Pew Internet and American Life project (Madden and Fox 2006). This analysis, which also attracted more than 140 citations, does not place the Wayback Machine at the center of the article, as do the other two; rather, it uses the Wayback Machine as a data source to analyze hits on MySpace versus Geocities. This approach is consistent with our use of the Wayback Machine as a data source rather than a web-based application. In the archive and library category, “Social Bookmarking Tools” by Hammond et al. was published in 2005 as a review article about link management approaches; the article attracted more than 170 citations and one of the approaches it profiled was an alternative to the Wayback Machine (Hammond, Hannay, Lund, & Scott, 2005). The most highly cited article in the legal area (receiving more than 280 Google Scholar citations) is a symposium lecture by Siegel (2006) on the defeat of the Equal Rights Amendment (ERA), which included a reference to a key letter opposing the ERA only available in the Wayback Machine.

Social science fields are also represented: 9% of identified journal articles are in various social science subfields such as technology management (2%), bibliometrics (1%), and business (1%). Several of the most highly cited works in this field are in the bibliometrics domain. Attracting 240 citations is the article titled “Webometrics” by Thelwall, Vaughan, and Björneborn (2005). The article describes various aspects of the web that can be used for quantitative analysis; relevant to this paper is the article’s pointing out that the Wayback Machine can serve as a useful source for developing time series data from past websites. Limitations of the search capability of the archive also were raised. Vaughan and Thelwall (2003) additionally produced “Scholarly Use of the Web: What Are the Key Inducers of Links to Journal Websites,” which, with 168 citations, demonstrated the relationship between links to journal websites and Journal Impact Factor ratings. The authors use the Wayback Machine as a data source for obtaining information about the age of a journal’s website. Carpenter (2007), in an international studies social science paper with roughly 150 citations, examined how transnational organizations develop advocacy agendas through a case study of interest in protection of children and girls in wars. In this research, the Wayback Machine was used to date the emergence of this topic on an archived web portal. Yadav, Prabhu, and Chandy (2007), in a study that has attracted nearly 130 citations, use the Wayback Machine as a data source to measure the breadth of internet banking deployment for analyzing the role of the chief executive officer (CEO) in encouraging innovation in banking. The authors use the Wayback Machine to count Java applets, internet business banking, internet tax filing, internet brokerage, and mobile banking applications associated with 176 banks.

These articles suggest that there is emerging but significant use of the Wayback Machine in multiple fields, including in social science research (Table 1). In some cases the Wayback Machine is used to locate a critical document and its first appearance or date. In other cases, the Wayback Machine has been utilized as a source of data for quantitative analysis. Our paper fits within this latter category, contributing to existing works by identifying and describing a process for retrieving and analyzing Wayback Machine website information for quantitative social science research.

The Webscraping Process

In this section, we describe our methodological approach to webscraping using the Wayback Machine. We illustrate this process using results from an analysis of U.S.-based SMEs in green goods industries. Green goods industries include the manufacturing of (a) renewable energy systems; (b) environmental remediation, recycling, or treatment equipment; and (c) alternative fuel vehicles, energy conservation technologies, and other carbon-reducing goods (for details about this definition, see Shapira et al., 2014). We abbreviate companies manufacturing in these industries as green goods companies (GGC). The objective in our use of archived GGC websites is to develop innovation activity indicators from historic company websites archived in the Wayback Machine to investigate the growth of GGCs. Our overarching goal in this paper is to provide an understanding of the steps involved with using the Wayback Machine in social science research based on our experience in this project. To this end, we walk through the six key steps: (a) sampling, (b) organizing and defining the boundaries of the web crawl, (c) crawling, (d) website variable operationalization, (e) integration with other data sources, and (f) analysis.

Step 1. Sampling

In the business sector, websites present a broad array of information about companies, products, and services, and thus offer promising opportunities for webscraping on a range of questions. Business website use has become widespread, although there are variations by types of firms. For example, in a companion analysis by Gök, Waterworth, and Shapira (2015) of GGCs in the United Kingdom, the
authors observe that just under three-quarters of U.K. companies with at least one employee report that they maintain a website. The percentage with a website rises to 85% for companies with 10 or more employees in all economic sectors in the United Kingdom and 91% for manufacturing companies with 10 or more employees. Yet, although larger firms typically maintain websites (some of which can be very extensive), many smaller firms also present a significant amount of information about their company in part because they use their website as a tool for selling services or attracting investment. In some cases, the amount of information on smaller firm websites is more extensive than one would see in websites of large corporations. Although it is true that there are smaller firms who may present less information, as well as some small firms not having their own online presence, the websites of those SMEs that are online can reveal indications of innovation activity. However, it should be noted that sometimes there are gaps in even these online firms’ web presence. Arora, Youtie, Shapira, Gao, and Ma (2013) report that some high-tech, small firm websites may “go dark” to protect their nascent and evolving intellectual property, especially when in preproduct stages. Their websites can remain online but with very limited content.

In order to use web scraping for data extraction, the analyst should first make sure the sample meets a few pre-requirements for web data. These pre-requirements include:

1. Do the organizations in the sample have websites?
2. Do their websites provide information to answer the analyst’s research questions?
3. Does the Wayback Machine capture the relevant historical websites?

We use the GGC analysis as the main example to illustrate how to build these pre-requirements into the sampling process. The research focus of the GGC analysis was on the innovation activities of U.S. GGCs and how these activities may or may not be associated with company growth. The basic requirement of our sampling approach was that these SMEs had websites. First, we found that the proportion of the population of GGCs with websites was relatively high, at least in the United States, particularly for those companies that have employees. In our GGC analysis, not all firms with current websites also had websites archived in the Wayback Machine, thus necessitating (at least) two rounds of formative sampling. In the first round, of the roughly 2,500 U.S.-based SMEs manufacturing in green goods industries, approximately 700 had current websites. Of these 700, only 300 both had manufacturing in some green goods area and had current and at least one archived website (see Shapira et al. [2014] for how “green goods” was operationalized). We subsequently found that nine of these companies had gone out of business. This gave us 291 GGCs in our database, although subsequent efforts to integrate with other data sources (specifically Dun & Bradstreet’s Million Dollar database which provided us with sales and employment growth figures), led us to discard 19 additional GGCs that had missing data points in the Dun & Bradstreet database.

Second, we found that current and archived websites of these SMEs had a diversity of information about company innovation activities and business approaches. Information about products, technologies, company news, and other business-related communications was much more readily available at the level of the SME than in large corporate websites, the latter of which tend to focus on financial reporting, corporate responsibility reporting, and broad market segments. In our GGC analysis, we found that nearly all the U.S. GGCs in our sample had a detailed product page or set of pages and two-thirds had information about their research and development activities (Arora et al., 2014).

### TABLE 1. Highly cited articles referencing the Wayback Machine. Papers referring to the Wayback Machine with 125 or more citations as of May 2014 in Google Scholar.

| Authors and year | Title                                                                 | Google Scholar citations (as of May 2014) | Category                   |
|------------------|----------------------------------------------------------------------|------------------------------------------|----------------------------|
| Resnik and Smith (2003) | The web as a parallel corpus                                      | 443                                     | Information technology     |
| Siegel (2006)     | Constitutional culture, social movement conflict and               | 283                                     | Legal                      |
|                  | constitutional change: The case of the de facto era.               |                                         |                            |
|                  | 2005–06 Brennan center symposium lecture                           |                                         |                            |
| Stein (2010)      | The case for cloud computing in genome informatics                  | 247                                     | Medical, information technology |
| Thelwall et al. (2005) | Webometrics                                                      | 240                                     | Social science—bibliometrics|
| Daughton (2004)   | Non-regulated water contaminants: emerging research                | 208                                     | Medical                    |
| Hammond et al. (2005) | Social bookmarking tools (I) a general review                    | 172                                     | Information technology     |
| Vaughan and Thelwall (2003) | Scholarly use of the web: what are the key inducers of links to      | 168                                     | Social science—bibliometrics|
|                  | journal web sites?                                                 |                                         |                            |
| Carpenter (2007)  | Setting the advocacy agenda: Theorizing issue emergence and        | 148                                     | Social science—international studies |
|                  | nonemergence in transnational advocacy networks                   |                                         |                            |
| Baker et al. (2006) | A fresh look at the reliability of long-term digital storage      | 143                                     | Information technology     |
| Madden and Fox (2006) | Riding the waves of “Web 2.0”                                    | 141                                     | Information technology, social science |
| Yadav et al. (2007) | Managing the future: CEO attention and innovation outcomes        | 128                                     | Social science—business    |
A third requirement concerns whether the Wayback Machine captured historical company website information. The Wayback Machine crawls websites in recurring yet not necessarily predictable, periodic intervals. For example, although some sites are crawled every month or so, others may only be revisited once a year or every other year. The Wayback Machine archives only publicly accessible pages, and pages protected by passwords or “do not crawl” exclusions (e.g., robots.txt files that disallow access) are not archived, nor are pages with embedded dynamic content (e.g., as enabled by JavaScript). In addition, pages not linked to another page (i.e., orphaned pages) also are not likely to be archived in the Wayback Machine. 2

The Wayback Machine says that it can take 6 or more months (and sometimes longer, up to 24 months) for archived pages to appear. 3 We found variations in the frequency of Wayback visiting any particular page, as well as the depth of the crawl given a “seed” page. In our GGC analysis, for example, one GGC had 34 pages in 2008, two pages in 2009, and six pages in 2010. Our experience suggests that some pages (e.g., home pages or whole websites associated with highly visible organizations) may be crawled more often than other pages, and there is no readily available explanation describing the variance in capture-rates between one page (or site) and another. Therefore, although in general the Wayback Machine offers a way to generate panel data sets by year with a site or corresponding firm as the unit of analysis, missing data are a significant challenge as firm data vary from year to year. We worked around this limitation by combining observations to produce longitudinal data pooled across multiple years. One of the studies coming out of our GGC analysis used this technique. Li, Arora, Youtie, and Shapira (2014) collected 4 years of website data (2008–2011) from the Wayback Machine for the 300 GGCs. But in order to construct a balanced data set, the researchers had to use two aggregate time period, 2008–09 and 2010–11, instead of each year.

Step 2. Organizing and Identifying the Boundaries of Web Crawl

Web scraping begins with identifying the boundaries of a collection of webpages, either current or archived, which is going to be searched and downloaded (“web crawling”) later. Although this article focuses on archived websites in the Wayback Machine, the step is also applicable to current webpages. This step needs to address two issues:

1. What kind of information presented on webpages need to be crawled?
2. Which webpages on an organization’s website should be included in the collection?

The immediate issue is determining what kind of information on the webpages can be captured. The casual user may personally view a website and look for common cues, such as design elements, text and graphics, menus (navigation), or “about us” pages. Yet capturing this information in web crawling is much more complicated. First, there is a great deal of variance in the way websites are designed and specified. For example, to what extent is Flash (a graphics and animation software platform) present, either in the whole site or just the menus, multiple languages, and sub-domains? Most webcrawling software cannot extract dynamic content in Flash platforms; thus, only limited data are able to be extracted from Flash-based websites by crawlers because the content is not in HTML (or variants thereof), rendering extracted data from these websites as incomplete representations of the original. Flash-based links between pages are also a problem in this regard, making links between pages not readily traversable. Second, the analyst has to rely on computer software to process the searching and downloading of a large number of webpages. While the capabilities of computer software have been growing with the adoption of increasingly powerful processors, there are still serious limitations on how many pages can be crawled. Third, although HTML pages in multiple languages are not usually a technical problem, some crawlers cannot handle a multitude of text encoding types (e.g., Unicode), so that multilanguage capability needs to be checked. In the GGC analysis, we employed IBM Content Analytics for web crawling (along with some custom Java code). IBM Content Analytics is a Linux-based software package with the capabilities to process HTML webpages as well as Adobe Portable Document Format (PDF) and Microsoft Word documents. The capabilities of other crawler software may vary.

The next step is to define the scope of a target website’s webpages to be crawled. Our general approach was to use a set of seed Uniform Resource Locators (URLs), that is, web addresses, and a number of corresponding domains to indicate which links will be traversed and which links are out of scope. The domains essentially act as boundary constraints in the crawl.

By way of explaining our approach to organizing and identifying the boundary conditions, we began by assuming D domains and S seeds, where D = S. The process for defining S and D is uncomplicated for “current” websites. For instance, to crawl “example123.com” as it currently appears on the World Wide Web (WWW), we set S = {http://example123.com} and D = {example123.com}.

Crawling the Wayback Machine complicates the process because the specification of S and D must include Wayback Machine URLs. 4 To illustrate, we first identified N seeds; next we created a domain pattern that captures everything on that site in a given time period. For this example, we retrieved pages in 2004 using two auto-generated seeds:

---

1https://archive.org/about/faqs.php (accessed December 18, 2014).
2https://archive.org/about/faqs.php (accessed December 18, 2014).
3https://archive.org/about/faqs.php (accessed December 18, 2014).
4The exact syntax will likely depend on the crawler.
The complications came when one of our GGCs either went public or entered the international market. In the case of going public, we saw a new investor relations subdomain (e.g., ir.ggcexample123.com). When the firm expanded to international markets, we saw websites that supported multiple languages in the form of a subdomain for each language (e.g., fr.ggcexample123.com) or as a “get” parameter specified toward the end of the URL (e.g., ggcexample123.com?lang=fr). To implement these additional crawling rules, we added to sets A and O as needed, for example:

\[ O = \{\text{https://web.archive.org/* fr.ggcexample123.com*}, \text{https://web.archive.org/*ggcexample123.com*lang=fr}\}. \]

In crawling the SME websites in the GGC analysis, we found that there were two specific concerns about company websites that required attention. First, three of the SME sites in our data set experienced a relatively uncommon, although still critical, boundary issue concerning subsidiaries. This situation occurred particularly with small firms that had grown rapidly into medium-size enterprises. In such cases, the firm over time diversified into multiple brands or divisions under a holding company. Each subsidiary was associated with a unique domain name in at least one case. As a result, each of these domains had to be crawled separately, and furthermore, once collected, the website information had to be re-aggregated and combined at the holding company level.

Second, one of the GGC websites of a battery maker contained hyperlink loopback mechanisms stemming from its extremely large and nested e-commerce pages. Although the page content will not substantively change, the URLs will, thus potentially resulting in continuous ad infinitum crawling. Therefore, in addition to specifying which domains to crawl, we had to specify which URLs to avoid, that is, avoiding certain URLs characterized by a particular regular expression or limiting the crawling depth to a certain number of levels. Crawling depth can be defined as follows: if a website can be traversed in an inverted tree-like structure, with the seed as the first page, then depth denotes the shortest path distance between the seed and any other hyperlinked URL. For example, we limited crawling of sites with very deep structures at 10 to avoid endless crawling.

In sum, we found that organizing and defining the boundaries of the web crawl may seem simple, but in fact it is a
time-consuming process requiring meticulous analysis of coverage issues on a year-by-year basis. Crawling unnecessary pages wastes valuable computing resources and time, but not crawling pages results in measurement errors with values differing unsystematically for each firm across the time periods under investigation. Thus, we took particular care in defining the boundaries of the web crawl because we understood that the integrity of the final data set and validity of the findings and interpretations depend on the integrity of the collections and boundary conditions.

**Step 3. Crawling**

The aim of the crawling process is to locate and store relevant archived websites defined by the boundary conditions described in Step 2. Given the large amount of repetitive work of searching and downloading involved in crawling, this step is often partly or fully automated with the aid of computer software. The automated crawling process has to be able to identify germane archived websites and omit nongermane sites. Relevant information on these sites is then stored for subsequent analysis. Issues involved in the crawling process include determining the type of information stored, the amount of crawling that can or should be done, and the need for recrawling to capture missing data.

In technical terms, crawling might be seen as an information retrieval process that requires a set of non-null seeds $S$, a set of non-null domains $D$, and a potentially null set of specific conditions $A$ and $O$ to “include” or “avoid” certain URLs (Figure 2). These parameters provide a minimum level of articulation of the crawler’s boundary conditions. From there, most crawlers allow for a playback mechanism that traverses the entire website structure and stores text and possibly other data such as images or PDF content. Many crawlers also store the hyperlink structure for later reconstruction.

Crawling ideally concludes when all pages in a domain have been “seen” by the crawler. To determine whether this is the case, one may turn to summary counts to discern whether at least a few (e.g., 3–5) pages have been crawled in a given time period: If no pages or only 1–2 have been crawled, there may be a problem with the crawling parameters $S$, $D$, $A$, or $O$. Alternatively, one may compare a site’s page counts in one time period vis-à-vis its counts in the next time period. However, as indicated above, coverage on the Wayback Machine is sometimes incomplete in a given year, so it may not be uncommon to see uneven representation from one year to the next. Bearing these concerns in mind, we manually verified missing observations for GGCs with very few pages. Of course, collecting data longitudinally in real time, that is, as the months or years pass by, is an acceptable alternative to using Wayback with the obvious limitation of adding additional time to the data collection effort.

The archived websites being crawled may be stored in “collections,” or units of raw data consisting of content from multiple websites that can be used for subsequent analysis. Depending on the quality of the data, a given website in a specific time period might have to be crawled more than once (“recrawl”). Many reasons may necessitate recrawling. These include misspecified boundary conditions, unexpected changes in website structures, or simply issues in the manner in which the Wayback Machine stores its archives. Recrawling can be targeted to specific domains or collections depending on the crawler technology and chosen collection configuration. To mitigate the risk of recrawling a large collection of crawled data (which is often a very time-consuming process), we implemented precautions by dividing the target websites and storing crawling results in multiple collections. In our experience with the GGC analysis, we allocated to each collection ∼50–150 domains (arrived at through trial and error) to accommodate the size of the collections that varied considerably by year. Depending on machine capabilities, more than 150 domains per collection may take too long to recrawl in case of errors in the crawling process, and may add unnecessary complexities to later data exporting. Too few domains per collection, on the other hand, may make data management too complicated, particularly in the data export process. We determined that 50–150 domains per collection achieved an appropriate balance of accuracy and economy. It should be noted that the Wayback Machine allows access to its archived content for scholarship and research purposes only, and that use of any part of the content is limited to noncommercial use.6

In terms of crawler software, we primarily used IBM Content Analytics in the GGC analysis.7 IBM Content Analytics offers options available at no cost to IBM Academic Initiative members (which can be joined by academic researchers without charge). This is attractive to our research group. Numerous other crawlers are available, albeit some with hefty licensing fees. Before choosing a crawler, we suggest reviewing its technological, crawling coverage, and human-led intervention/testing capabilities.

**Step 4. Website Variable Operationalization**

Although the crawled results are often sufficient for analysis in computer and information science, there is one additional but critical step for social science analysis—the conversion of web data into social science variables for analysis. Website content is essentially unstructured data, with significant noise alongside key information of interest. The issue thus becomes how to filter the noise from the information for representation as social science variables. There are a number of newly developed methods, including keyword-based and natural language processing approaches, advanced unstructured text-mining algorithms, and network-based analytic methods that facilitate the process of separating focal content from noise and facilitating the process of extracting meaningful insights from unstructured text.

---

6Internet Archive’s Terms of Use, Privacy Policy, and Copyright Policy. https://archive.org/about/terms.php (accessed December 18, 2014).
7Subsequently rebranded as Watson Content Analytics. see: http://www-03.ibm.com/software/products/en/watson-content-analytics (accessed January 24, 2015).
Perhaps the most straightforward approach for operationalizing social science variables from website data is to construct variables by counting relevant words, keywords, and phrases. This is primarily what we used in our GGC analysis. We used these counts to represent stocks or flows. Similar to their economic meanings, stocks refer to a cross-sectional snapshot of a variable’s value, while flows are the difference in stocks from one period to the next. We employed simple keyword extraction to relate core concepts to operationalized variables through counting mentions of predetermined terms in the website content.

The effectiveness of keyword extraction and counting is, of course, dependent on the structure of the target website, which is partly a function of the breadth and depth of content. As well as the number of words on each page. For example, we found that some online media (and to a lesser extent firm websites) will segment content across multiple pages. We thus decided that it would be useful to normalize keyword and phrase counts by the number of pages on a site, the number of words on a site, or the average number of words per page on a site. This normalization enables comparison of variable values across websites. We had to pay special attention, however, to a weakness of keyword extraction, which is that certain terms were not used or interpreted in the same way across multiple websites. Gök et al. (2015) report on these differences in a research and development (R&D) context, highlighting how the simple use of the term “development” attracts a number of false-positive mentions (for instance, associated with property development) in approximately 300 U.K. green manufacturer websites. They conclude that different term and phrase combinations can produce significant differences in aggregate results, and that fine tuning and contextualizing search terms can improve reliability.

Related to keyword extraction is Named Entity Recognition (NER) software that identifies specific patterns of unstructured text in three categories: people, places, and organizations. NER engines work to varying levels of effectiveness, depending on the algorithmic implementation and input data. A widely used, free package is available from Stanford’s Natural Language Processing Group (Finkel, Grenager, & Manning, 2005). Included in Stanford’s NER are three models with three, four, and seven respective classes trained on two data sets: the three-class model corresponds to the aforementioned identification of people, places, and organizations, while the four-class model adds a miscellaneous label. The seven class model tags locations, organizations, person names, times, money and currency, percentages, and dates. Another popular tool for NER, and indeed many other unstructured text processing tasks, is GATE (General Architecture for Text Engineering) (Cunningham, Maynard, & Bontcheva, 2011; Cunningham, Tablan, Roberts, & Bontcheva, 2013).

To adequately use keyword extraction, the search terms are ideally derived from social science concepts. While NER can be seen as a special case of keyword extraction, there are no standard ways of mapping the tagged content from NER to social science concepts. In our project, we carefully examine research context and relevant literature before linking the data with specific concepts. Here, we offer two examples of assigning NER-based variables to concepts in innovation research. In the innovation and network literature, for example, increases in linkages between organizations in certain industries suggest a resource advantage (Ahuja, 2000; Borgatti & Foster, 2003; Brass, Galaskiewicz, Greve, & Tsai, 2004). Therefore, examining changes in the number of organizations mentioned on a firm’s website from one time period to the next may serve as a proxy for network-based resource advantage. Alternatively, person names may be used to identify specific human capital at a firm, or changes in strategy over time as proxied by personnel changes. Naturally, identifying who is an employee on a website versus a reference to an external individual requires a separate analytic filter. We reflect on these challenges in more detail in a discussion of construct validity below.

In the GGC analysis, we mainly used keyword searching and NER to construct variables from website data to measure SME innovation and entrepreneurial activities. The list of search keywords is derived from a literature review, workshops with specialists, and two prior manual-based analyses of SMEs’ current and archived websites in nanotechnology domains (Arora et al., 2013; Youtie, Hicks, Shapira, & Horsley, 2012). For example, entrepreneurship scholars emphasize the importance of venture capital and other equity investment as one way for small startups to overcome resource scarcity and the inability to secure collateral-based lending or debt financing (Amatucci & Sohl, 2007; Auerswald & Bozkaya, 2008). To capture this feature of SME activities that are essential to innovation, we included keywords in this variable class such as venture capital, private equity, private placement, and seed funding.

We used the NER method as provided through IBM Content Analytics to capture locational variables and wrote a Java program to clean and aggregate place-based names. For example, we constructed a “local” variable to measure the GGC’s interaction with local organizations and entities; according to economic geography literature, one would expect that proximity might bring additional benefits for social interactions and innovative activities. Yet for each SME, “local” has a contextual meaning related to its geographical location. We used the Office of Management and Budget (OMB) definition of combined statistical areas/metropolitan statistical areas (CSA/MSA) to define an SME’s “local” region. For example, for an SME located in the St. Louis area, “local” could be “St. Louis,” “St. Louis, MO,” “St. Charles, MO,” “Madison,” “Granite City, IL,” or any location in the St. Louis MSA. These locational terms were fed into the NER software function in IBM Content Analytics and cleaned through our Java program to group cities in the same CSA/MSA and count them to generate the variable “local” for this specific SME. This process had to be repeated for each SME.

In addition to keyword and phrase searching and NER to extract specific information, there are methods to take advantage of website data as a whole for additional analysis.
A research area in computer science receiving significant attention is topic modeling, and particularly variations of latent Dirichlet allocation (LDA) (Blei, 2012; Blei, Ng, & Jordan, 2003). LDA approaches unstructured text as a “bag of words” where latent topics consist of a probability distribution over observed words, and these topics in turn are assigned varying probabilities to documents in the corpus. An alternative method is “term clumping,” in which standard types of noise such as stop words, common scientific and technical terms, stand-alone numbers, and the like are culled out; phrases with similar words are combined, and the results are analyzed using dimensional reduction techniques such as principal components analysis (Porter, Newman, & Newman, 2012). In both approaches, the resulting topics or factors can be turned into variables to enable analysis of, for example, how the websites of organizations have changed their topical orientation. We used LDA to assess the extent to which topics on a GGC’s website changed between two different periods (see step below).

Steps 5 and 6. Integration With Other Data Sources and Analysis

The data set obtained through webscraping can be analyzed using most social science analytical software programs. These methods include basic descriptive reporting, linear (or nonlinear) modeling, clustering and dimension reduction, and case study approaches. Although it is usually not an issue in a conventional data set, we found that we had to pay special attention to which social science concepts the webscraped variables were measuring. Specific examples are illustrated below.

In the first example, variables derived from keyword extraction of a GGC’s archived websites were used to measure a firm’s relationship with external innovation partners (Li et al., 2014). These measures of relationship are usually hard to obtain except either from surveys and interviews or proxied by coauthorship in publications and patents. However, we saw that, at least for GGCs, the incidence of firms that had published or patented in green goods industries was less than 20%, even though more than 60% of U.S. GGCs engaged in R&D based on their archived websites (see also Gök et al., 2015). The variables derived from webscraping thus provided us with alternative, and in most cases, more complete measures. In a panel model, the impact of relationships with three types of external partners—university, government, and industry—was estimated. We determined in the course of this analysis that webscraping was useful to gauge university and industry relationships, but not government relationships, so we integrated data from the sam.gov website into our webscraped data set to obtain information about which U.S. GGCs had registered to contract with the federal government.

In the other study, the authors derived variables obtained through topic modeling techniques using the same Wayback Machine data set. Arora et al. (2014) trained LDA on a sample set of firm webpage documents in the 2008–2009 time period, and then inferred the amount of topical change for a firm in this 2008–2009 time period vis-à-vis that same firm’s topical content in the 2010–2011 time period: the greater the change in topical content, the more dissimilar the firm’s topical emphasis is between the first and second time period. The authors used this information distance measure as a proxy for strategic pivoting, or the propensity for small technology firms to change their strategic orientation as a result of searching for a market niche for their products and services.

In both studies, variables derived from webscraping were used to measure social science concepts that are difficult or impossible to obtain with conventional methods. To model the effects of innovation indicators derived from webscraped data on firm outcomes such as sales or employment growth, we combined webscraped-based data sets with conventional data sources with economic and social information. This data integration required careful matching of the GGCs in our webscraped data with the names of companies in these secondary data sets. In our case, we matched webscraped data with companies in Dun & Bradstreet’s Million Dollar database. We used a combination of these sources’ unique identifiers along with the firms’ names, albeit sometimes we came across firms that had undergone a name change, so we had to do further checking to ensure that we were matching the correct company’s sales and employment information. As a result of this integration process, we had to drop 19 GGCs from our analysis because, although they had websites, they had missing information about sales or employment for a particular year of interest in our analysis.

Conclusion and Discussion

In the previous section, we reviewed the major steps of conceptual and technical implementation of utilizing the Wayback Machine as a data source for social science research. Drawing on the experience in the analysis of 300 U.S.-based GGCs, we highlight the challenges from webscraping and our suggested approaches to address them. Here we summarize three lessons learned with the intention of providing the broader research community with helpful and time-saving insights.

First, we found that capturing website data from the Wayback Machine is more difficult to scale than expected. In principle, the analyst sets the seed and domain of the crawler and lets the technology do its work. In practice we faced issues that varied on a site-by-site, year-by-year basis and that required manual attention. The flexibility of the web that has made it such a boon as a general purpose innovation also makes it difficult to apply a systematic approach to. For example, being able to diffuse news over the web—e.g., by copying content on a single website—vastly increases the amount of available data. The researcher must decide whether duplicate data matters, and if so, how to measure it. The presence of outliers, reliability in the method, and construct validity are essential to valid
interpetations of findings. Because the web (and archive.org) is a disordered place, any research using website data should account for this in terms of design, time, and research effort.

Second, the notion of construct validity deserves special attention because webscraping for the social sciences inherently tries to make sense out of unstructured text. Unlike many structured databases, sources of unstructured text do not contain metadata, and therefore attempting to link webscraped variables to social science concepts requires careful consideration of research questions. Again returning to the context of the GGC analysis, we consider three examples to illustrate issues related to construct and internal validity.

• University linkages. Much of the innovation literature posits a direct relationship between positive business outcomes and university-facilitated entrepreneurship (Hess & Rothenberg, 2012; Roithammer, Agung, & Jiang, 2007). For example, the university scientist can offer important intellectual capital to a biotech startup, or a university’s business accelerator can provide resources and connections that would not otherwise be available to non-incubated firms. In the GGC analysis, we counted the number of university mentions on a website as a proxy for university linkages. At the same time, we acknowledge that the exact nature of a linkage may differ from one webpage to the next. By comparing quantitative measures to the raw text found on select high-tech small firm websites, we found three types of university references: spin-outs, licensing of intellectual property, and institutions attended by senior management for academic training. Because of these differences, counting university mentions at large included a variety of formal and informal linkages. While we did not need to disambiguate the type of university reference for the GGC work (all three types of university references were relevant for the GGC analysis), other social science research questions may require more manual checking. For example, if a hypothetical social science study were to investigate a research question about companies’ explicit information exchanges with universities, that study would likely exclude pages referring to company management teams’ older university degrees, as such older degrees would be unlikely to generate recent university-industry information exchanges.

• Locations. The economic geography literature argues that location matters: embeddedness in a robust industrial district, milieu, or regional innovation system conveys certain resources and advantages that more geographically dispersed (and less embedded) firms are not as likely to experience (Cooke, 2001; Doreux & Parto, 2005; Simmie, 2005). Alternatively, other research suggests that start-up firms operating in niche markets may go directly to the international market, that is, become “born global,” to find customers or innovation partners (Knight & Cavusgil, 2004; Oviatt & McDougall, 1994, 2005). Our GGC work proxied geographical embeddedness at the local, national, and international levels (a) by identifying the small firm’s home CSA/MSA; (b) via NER, by counting all geographic locations that are within the local region, in the home country but outside of the local region; and (c) then counting international locations (cities and nations). We used this process to proxy the geographic orientation of each GGC. We excluded the primary location name(s) of the company under analysis, and within this limitation, there are many examples that city and state references are valid indicators of market orientation. For instance, one GGC located in Georgia, USA, highlighted on its website information about its solar panel work in the neighboring state of Tennessee. Likewise, a university spinoff listed on its website several East Asian countries where it distributes its products. Of course, the difference between sourcing in one place and exporting to that same locale can be a critical distinction in innovation, strategy, and operations research, although this distinction was not important to our GGC analysis. Still, because of these reasons, and within the limitations discussed here—which can require some manual checking—we believe that mentions of university and location generally have construct validity.

• Endogenous variables. The way in which website variables are constructed may give rise to additional internal validity concerns. For example, website variables are often endogenous regressors when employed in econometrics models to predict business performance such as sales or employment growth, all of which are related to company size and resource availability. Since websites consume corporate resources to deploy, revise, and maintain, companies with larger sizes, faster growth, and/or more resources may be more likely to update and enlarge their websites. This leads to bidirectional causality between explanatory website variables and predicted outcomes; higher values of web indicators may predict better performance, but better performance may also predict higher values of those same web variables. To address endogeneity as a threat to internal validity, we employed in the GGC analysis established econometric approaches such as fixed-effects models, which drop unobserved time-invariant differences across the panel, or instrumental variable approaches, which attempt to isolate and remove the spurious effects of endogenous regressors. In our research on the effect of external relationships measured by website variables on the growth of U.S. green goods SMEs, we constructed a fixed-effect panel model using additional exogenous regressors obtained from secondary data sources (i.e., Dun & Bradstreet and the U.S. Bureau of Economic Analysis) to improve model specification (Li et al., 2014). We also included Hausman Taylor estimators to obtain consistent estimation results when specifying a mix of time-variant and time-invariant regressors in the fixed-effect model.

Third, we contend that using websites as a social science data source requires a strong set of multidisciplinary skills. In related emerging fields, such as social network analysis and computational social science, scholars remark that gaps often exist between researchers’ skills and the empirical setting (Borgatti, Everett, & Johnson, 2013; Lazer et al., 2009). One approach is to assemble a multidisciplinary team, rely on individual specialized training, and then attempt to combine the outputs into an interdisciplinary project. Based on our experience, a second approach may be more effective: find interdisciplinary people that are trained in social science, data management, computer science, etc., to shepherd the project. This latter approach requires less oversight because it does not need to deal with diverse discipline-specific ways of viewing the research setting.
One such skill area distant to many social scientists is the realm of software development and statistical analysis using unstructured text. We have found that each stage of the data pipeline required a set of interfacing tools. For example, to extract the data and operationalize keyword counts, we employed bibliometric applications in the VantagePoint software or custom scripting code using Java. For more leading-edge work developed out of the computer science community, libraries could be used as an alternative to building custom code, although we did not use libraries in our analysis. In most cases, competency with programming is likely a necessary condition for the effective utilization of unstructured website text to explore social science research questions.

The ability to use software to automate the process of crawling, collecting, and analyzing information in the Wayback Machine enabled our project team to create and analyze a data set that was an order of magnitude larger than we were able to create through manual means alone in prior studies (Youtie et al., 2012). This ability to scale-up the size of the data set permitted us to create econometric models to test factors that were expected to be associated with growth in GGCs. However, achieving this scale does require a significant manual verification effort. The sampling step involved manual checking of the availability of a firm’s archived websites. Organizing the boundaries of the web crawl involved manual confirmation that the specification of the crawl indeed captured a firm’s archived website information to the greatest extent possible. Manual review of crawling indicated that one or more archived versions of a firm’s website had to be recrawled because of unexpected changes in the way the website information was stored in the Wayback Machine. Manual investigation of website variable operationalization indicated that a particular keyword may have attracted false positives, as in the example of the term “development,” which captured archived web information about property development rather than information about R&D, as was intended (Gök et al., 2015). Integrating other data sets involved manual inspection to make sure that a given firm was accurately matched in the archived and external data sets. Initial analyses identified outliers that gave cause for returning to the archived data set to manually check the accuracy of the archived information on which a variable’s values were based. In sum, optimal use of the Wayback Machine for social science analysis combines automated methods with manual verification at all steps in the process.

Websites offer great potential in researching social science phenomena. In this paper, we conveyed the utility of web scraping and analysis in the context of one specific research domain of interest, small innovative firms in green goods industries. We offered a six-step approach to facilitate replication of our processes by the website analysis and social science communities. We hope this work offers a helpful methodological resource to the broader social science community interested in but not yet comfortable with harnessing online data. Although the Wayback Machine is a valuable resource for scholars in all fields, we believe that social scientists in particular have yet to tap the full potential of the treasure-trove of data found online. Going forward, there are significant opportunities to incorporate additional analytical tools and statistical models while at the same time testing, refining, and reporting on the validity of measures and results. However, one should keep in mind that both automated and manual effort is needed to create high-quality information capable of providing these benefits and overcoming the limitations of the archived information.

Acknowledgments

This work was supported by the U.K. Economic and Social Research Council (grant number ES/J008303/1) and by the U.S. National Science Foundation through the Center for Nanotechnology in Society (Award No. 0531194 and No. 0937591).

References

Ahuja, G. (2000). Collaboration networks, structural holes, and innovation: A longitudinal study. Administrative Science Quarterly, 45(3), 425–455.
Amatucci, F.M., & Sohli, J.E. (2007). Business angels: Investment processes, outcomes and current trends. In A. Zacharakis & S. Spinelli (Eds.), Entrepreneurship: The engine of growth (vol. II, pp. 87–107). Westport, CT: Praeger Publishers.
Arora, S.K., Youtie, J., Shapiro, P., Gao, L., & Ma, T. (2013). Entry strategies in an emerging technology: A pilot web-based study of graphene firms. Scientometrics, 95(3), 1189–1207.
Arora, S.K., Youtie, J., Shapiro, P., & Li, Y. (2014). Does “Pivoting” matter for innovation? The usefulness of strategic business changes in analyzing growth in small and medium-sized green goods firms. Paper presented at the 23rd International Management of Technology Annual Conference 2014, Washington, D.C.
Auerwald, P.E. (2008). Financing entrepreneurship, vol. 12. Northampton, MA: Edward Elgar.
Auerwald, P.E., & Bozkaya, A., eds. (2008). Financing entrepreneurship. Northampton, MA: Edward Elgar.
Baker, M., Shah, M., Rosenthal, D.S., Roussopoulos, M., Maniatis, P., Giuli, T.J., & Bungale, P. (2006). A fresh look at the reliability of long-term digital storage. ACM SIGOPS Operating Systems Review, 40(4), 221–234.
Baruch, Y. (1999). Response rate in academic studies—A comparative analysis. Human Relations, 52(4), 421–438.
Blei, D.M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77–84.
Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.
Borgatti, S.P., & Foster, P.C. (2003). The network paradigm in organizational research: A review and typology. Journal of Management, 29(6), 991–1013.
Borgatti, S.P., Everett, M.G., & Johnson, J.C. (2013). Analyzing social networks (p. 304). London: SAGE.
Brass, D.J., Galaskiewicz, J., Greve, H.R., & Tsai, W. (2004). Taking stock of networks and organizations: A multilevel perspective. Academy of Management Journal, 47(6), 795–817.
Carpenter, R.C. (2007). Setting the advocacy agenda: Theorizing issue emergence and nonemergence in transnational advocacy networks. International Studies Quarterly, 51(1), 99–120.
Cooke, P. (2001). Regional innovation systems, clusters, and the knowledge economy. Industrial and Corporate Change, 10(4), 945–974.
