Journal of the Midwest Association for Information Systems

Volume 2025 | Issue 1 Article 2

Date: 01-31-2025

Organizing Big Web Data: The Tidy Qualitative Data Concept and Criteria

Vlad Krotov
Murray State University, vkrotov@murraystate.edu

Abstract

The vast repositories of qualitative data available on the web present academic and industry researchers with numerous opportunities for answering important research questions with more rigor, precision, and timeliness. Harnessing this data for the purpose of further analysis requires putting it into a tidy format. To assist researchers with organization of qualitative web data, this paper develops the concept of tidy qualitative data. A list of seven specific criteria is proposed to operationalize this definition and provide practical guidelines to researchers on how to prepare qualitative data from the web for further analysis. A short case study illustrating an application of these principles of tidy qualitative data is also provided.

Keywords: big data, web data, qualitative, tidy data, tidy qualitative data

DOI: 10.17705/3jmwa.000092

1. Introduction

For decades, social scientists regarded data as a valuable and rare treasure, often requiring either good fortune or extensive effort to collect (Munzert et al., 2015). Today, due to the increasing digitalization and virtualization of social processes, data are less of a treasure: there are vast oceans of data floating on the World Wide Web. In fact, the data currently available on the web is measured in zettabytes (1 ZB = 10^{21} bytes or 1 trillion gigabytes) (Cisco Systems, 2016). The recent COVID-19 pandemic changed the nature of work through automation and the digitalization of work processes previously thought to be confined to the physical realm (Hadidi & Klein, 2024). This digitalization of processes led to another powerful "explosion" in the volume of digital data available on the web (Krotov & Johnson, 2023). The text generation capabilities of modern generative AI (GAI) tools, such as ChatGPT, Microsoft Copilot, Google Gemini, and Jasper, make the generation of digital content for students, educators, and corporations easy and fast (Bansal et al., 2023). Thus, GAI is likely to lead to another powerful "explosion" in digital data on the web.

When people think about "big data," they often envision powerful *quantitative* tools and analysis techniques being used to store, manage, and analyze such data. But much of the data available on the web is comprised of semi-structured *qualitative* data in the form of web pages, blog posts, social media posts, customer reviews, etc. (Watson, 2014). This is due to the fact that humans communicate, for the most part, using languages and text and not numbers and formulas (and are likely to continue doing so in the foreseeable future). Thus, human communication and interaction are the main drivers behind this growth of qualitative data on the web.

This qualitative data presents researchers in social sciences with lucrative opportunities for studying numerous sociotechnical phenomena with more precision, rigor, and timeliness. The present challenge for many researchers is not where to find suitable data but rather how to retrieve and organize it in a way that can enable further analysis via a variety of analytical approaches (Krotov & Johnson, 2023; Krotov et al., 2020; Krotov & Tennyson, 2018; Krotov & Silva, 2018). In fact, approximately 80% of the time devoted to a research project is usually spent sourcing and preparing data for analysis (Wickham, 2014). Sourcing and organizing qualitative data are especially challenging due to its volume, variety, veracity (Goes, 2014), and great dependence on the socio-technical context within which the data is produced and collected (Klein & Myers, 1999). Given these challenges, harnessing this data requires a socio-technical approach (Krotov & Johnson, 2023).

The main contribution of this paper is the development of the "tidy qualitative data" concept. This concept is formulated to guide researchers on how to organize qualitative data sourced from the web. Tidy qualitative data is defined in this paper as qualitative data that is organized and stored in a way that preserves its context and makes it easy to analyze with a variety of qualitive techniques. This definition is supplemented with specific criteria or recommendations for making a "messy" qualitative dataset a "tidy" one. Putting qualitative data into a tidy format will help researchers improve the rigor of their analysis, make their research findings more reproducible, and increase the likelihood of their studies being further extended by other researchers (Peng, 2011).

2. Literature Review

2.1 The Challenges of Sourcing Web Data

While qualitative web data is extensive and often available to researchers free of charge, retrieving and organizing this data from the web is often a project on its own. Given the volume, variety, velocity, and veracity of qualitative data available on the web, the collection and organization of web data can hardly be done manually even if large teams of researchers participate in data collection. Instead, researchers usually rely on technology tools to automate at least some aspects of web data collection and organization. This somewhat new practice of using technology for collecting and organizing digital data from the web is called "web scraping" (Krotov & Johnson, 2023; Krotov et al., 2020; Krotov &Silva, 2018; Krotov & Tennyson, 2018). Web scraping is commonly used in information systems (IS) research and other fields of social science (e.g., Kearney & Liu, 2014; Triche & Walden, 2018; Vaast et al., 2017).

Web scraping is a complex, socio-technical process that includes three interrelated phases: website analysis, website crawling, and data organization (Krotov & Tennyson, 2018; Krotov & Silva 2018) (see Figure 1). Although technology is usually used in all three phases, these phases still require a certain degree of human supervision in case errors or setbacks arise (e.g., wrong data elements being selected from a web page or a web server where the page is hosted becoming unresponsive).

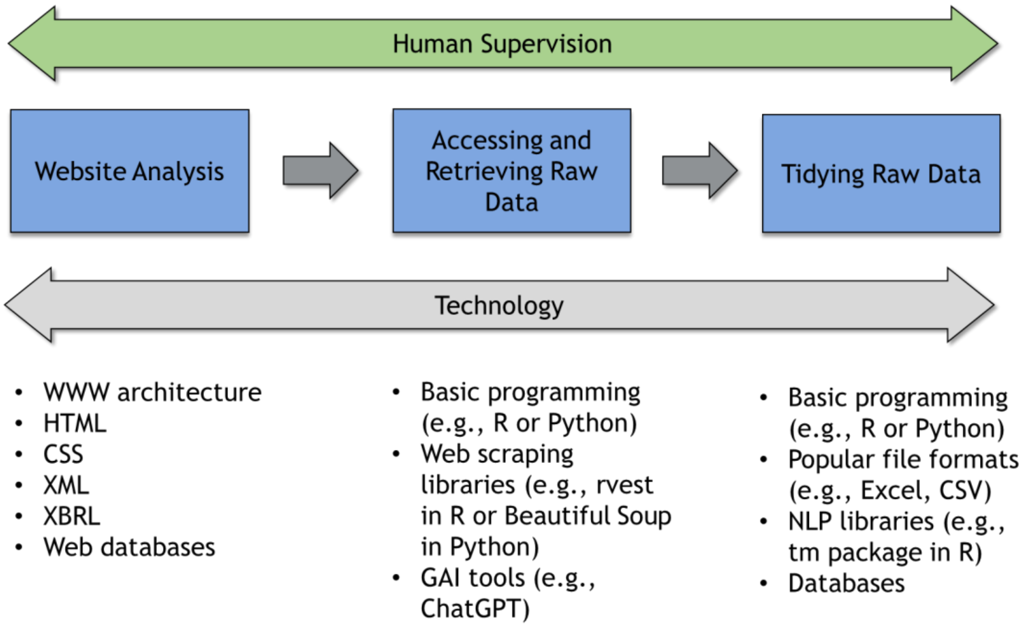

Human Supervision
Website Analysis
Accessing and Retrieving Raw Data
Tidying Raw Data
Technology
• WWW architecture
• HTML
• CSS
• XML
• XBRL
• Web databases
• Basic programming (e.g., R or Python)
• Web scraping libraries (e.g., rvest in R or Beautiful Soup in Python)
• GAI tools (e.g., ChatGPT)
• Basic programming (e.g., R or Python)
• Popular file formats (e.g., Excel, CSV)
• NLP libraries (e.g., tm package in R)
• Databases

Figure 1. Web Scraping
Note. Reprinted from Krotov and Silva, 2018

The purpose of website analysis is to understand how data are stored and organized on a website (or web repository, such as an online database) (Krotov & Johnson, 2023; Krotov et al., 2020; Krotov & Silva, 2018; Krotov & Tennyson, 2018). For this stage, one will need a basic awareness of the World Wide Web architecture and the type of information commonly found on the Internet. Furthermore, this phase requires an understanding of some of the most commonly used web technologies, such as HTML, CSS, XML, XBRL, JSON, and MySQL.

The web crawling phase of web scraping involves developing and running a script that automatically browses (or "crawls") the entire website or web repository in order to access and save the needed data (Krotov & Johnson, 2023; Krotov et al., 2020; Krotov & Silva, 2018; Krotov & Tennyson, 2018). Web crawling scripts are often developed using programming languages such as R and Python. These languages are popular for web scraping and data science because they include a number of libraries for crawling and parsing web data (e.g., rvest package in R or Beautiful Soup library in Python). A GAI tool like ChatGPT is often used to generate Python, R, and other web scraping scripts. AI can generate such scripts, but debugging and finetuning these scripts still require human attention in most instances.

After the necessary data are retrieved from a web repository, it needs to be cleaned, preprocessed, and organized in a way that enables further analysis of the data (Krotov & Johnson, 2023; Krotov et al., 2020; Krotov & Silva, 2018; Krotov & Tennyson, 2018). This phase is called tidying raw data. To save time, a programmatic approach may also be necessary due to the volume of data involved. There are libraries available for cleaning and organizing data in many programming languages. The tm package in R, for example, contains numerous functions for cleaning and preprocessing textual data, such as functions for removing white spaces and stemming the collected text.

Several software tools are available to help researchers automate most (if not all) of the steps in web scraping (e.g., import.io). Nevertheless, some level of human supervision is usually necessary to obtain clean, tidy data from the web. Oftentimes, "ready-made" web scraping tools select the wrong data elements from a web page. This usually happens due to the fact that numerous web standards (e.g., HTML, CSS, XML, XBRL) are often loosely interpreted and implemented by countless web developers from all over the world. Moreover, numerous networking errors may arise during the scraping process (e.g., an unresponsive web server). The author of this paper had a situation where his IP address was temporarily banned by the web server from which data was being retrieved (perhaps out of suspicion that this was a denial-of-service attack, as the script would send data requests to the server at a high rate). Because of that, the script had to be modified so that there was a one-second pause between the requests sent to the server. Also, a script running overnight to perform a large web-scraping task was once interrupted by a power outage in the building where the computer was hosted. As one can see, all these errors and setbacks require troubleshooting by a human. Thus, web scraping cannot be fully automated (Krotov & Silva 2018; Krotov & Tennyson 2018).

2.2 The Tidy Data Concept and Criteria

Regardless of which approaches or tools researchers choose to use to collect qualitative data from the web, the data need to be tidy in order to facilitate further analysis, improve the reproducibility of the findings, and help other researchers extend the results of the study by initiating new research projects based on the data (Peng, 2011). The tidy data concept was first proposed by Hadley Wickham (2014), a chief scientist at RStudio and a well-known contributor to the R language. According to Wickham (2014), "like families, tidy datasets are all alike, but every messy dataset is messy in its own way" (p. 2). Tidying a "messy" dataset involves structuring it in a way that facilitates further analysis. A dataset is considered to be tidy when it meets the following three criteria: "(1) each variable forms a column; (2) each observation forms a row; (3) each type of observational unit forms a table" (Wickham, 2014, p. 4). The three criteria are parallel to Codd's (1972) third normal form, albeit formulated in the language used by many statisticians and data scientists.

As shown in Figure 2, tidy data is an important part of the data science process. An organized, clean dataset is essential for the application of a variety of analytical techniques to the data so that theoretical generalizations can be drawn from the findings. Using tidy data, we can also build useful data products that help us better understand the world. It is also important to note that data sourcing is largely driven by research questions (or questions) about the world. Findings derived from research questions allow researchers to refine their research questions or formulate new ones; this often requires finding new datasets.

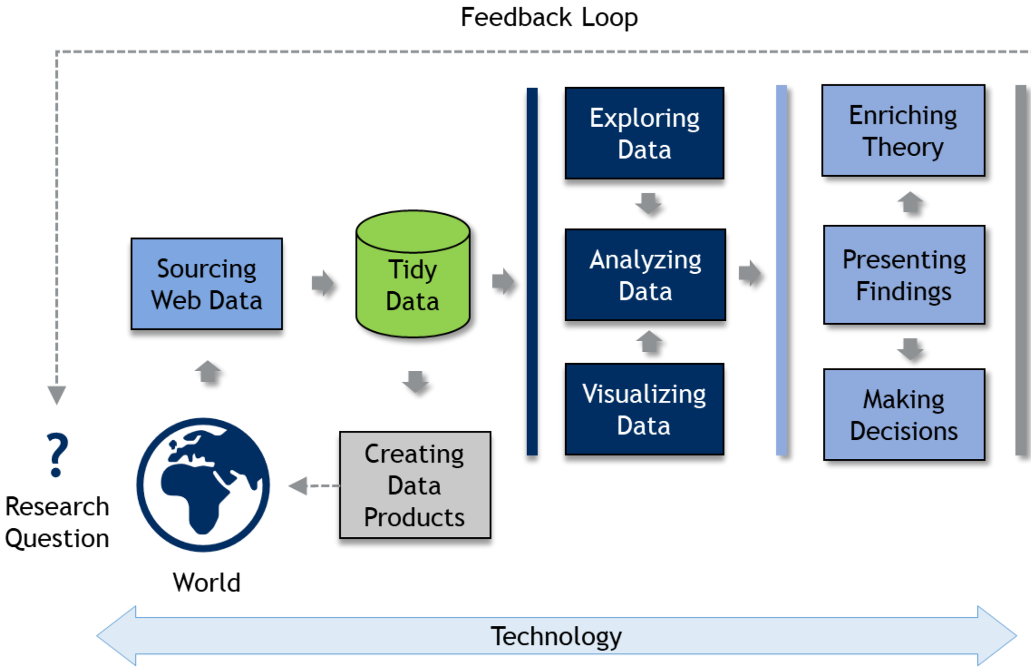


Figure 2. Tidy Data and Data Science
Note. Adapted from Schutt and O'Neil, 2013

It can be argued that Wickham (2014), a statistician by training, is implicitly dealing with quantitative data when formulating his tidy data criteria. While most of these principles are applicable and useful to qualitative researchers, it can also be argued that some elements or criteria central to the qualitative tradition of research are missing. For example, in qualitative research, the social context within which data is collected is extremely important and, thus, needs to be described and reflected upon (Klein & Myers, 1999; Walsham, 1995). This information is useful in interpreting the findings derived from this data by relating it back to the context where it was collected. Also, social scientists pursuing qualitative research projects do not deal with variables, but rather with units of observation (Eisenhardt, 1989; Yin, 1994). Thus, the definition of quantitative tidy data proposed by Wickham (2014) needs to be modified and expanded to fit with the qualitative research tradition in social sciences. Moreover, the definition needs to be expanded and operationalized with a list of specific, actionable requirements that can help one prepare a tidy qualitative dataset.

3. Tidy Qualitative Data

3.1 Definition

Drawing on the concept of tidy quantitative data proposed by Wickham (2014) as well as the qualitative tradition in information systems research (Eisenhardt, 1989; Yin, 1994), tidy qualitative data is defined as follows: qualitative data that is organized and stored in a way that preserves its context and makes it easy to analyze with a variety of qualitive techniques. This definition of tidy qualitative data translates the tidy data concept by Wickham (2014) into the language used by social scientists belonging to the qualitative research tradition and includes researchers who use automated, semiautomated, or manual tools for qualitative data analysis.

3.2 Tidy Qualitative Data Criteria

The concept of tidy qualitative data is further clarified and expanded by proposing a list of specific criteria that a tidy qualitative dataset retrieved from the web must adhere to. These criteria are important rules and considerations that qualitative researchers keep in mind when collecting, organizing, and analyzing qualitative data as part of their research projects.

Each criterion is listed below and supplemented with explanations and examples.

Criterion 1: The data needs to be stored in a format easily accessible by humans and computers.

Ideally, a researcher needs to select a format for storing qualitative data that is convenient for both humans and computers. Some examples of popular file formats readable by both humans and computers include Microsoft Excel, text, and comma separated values (CSV) files.

Criterion 2: The data is clean.

A clean qualitative dataset is free from such “impurities” as extra white spaces, unreadable characters, markup tags, and repeated or redundant elements. These “impurities” can stand in the way of gaining a better understanding of the meaning of this data, unnecessarily increase the size of the dataset, complicate further analysis, and impact the validity of findings derived from this data. Many of these “impurities” can be removed using natural language processing (NLP) tools and packages (e.g., the tm package in R). It should be noted that “clean” does not equate to “tidy,” as tidy data has to meet all of the criteria listed here.

Criterion 3: Data is organized according to the unit of observation.

Examples of units of observation used in qualitative research include: an email, a job post, a document, a social media post, etc. One can use a table to store information about many units of observation. Alternatively, a separate file can be used for each observation (if units of observation are large).

Criterion 4: Each observation should be clearly demarcated from other observations.

For example, if data is tabulated, then rows can be used to separate each observation. If observations are used as separate files, then the file structure provides the needed separation. In many cases, it is helpful to use a unique identifier (the so-called “primary key”) for each of the units of observation. In any case, the separation of the units of observation needs to be clear to both humans and computers to provide for a structured and meaningful analysis of the data.

Criterion 5: Metadata elements should be added to each unit of observation.

Each unit of observation should be supplemented with all relevant metadata elements (e.g., the date and time when the unit of observation was created, when it was retrieved, the source of the unit of observation). These metadata elements will aid in recreating the context within which the data was collected and retrieved and interpreting the findings derived from this data.

Criterion 6: The data and metadata elements should be labeled properly.

Each label used to describe a unit of observation or a metadata element should be brief, yet intuitive enough so that a researcher can quickly grasp its meaning.

Criterion 7: The dataset should be supplemented with a description file.

The main goal of the description file is to communicate the socio-technical context within which the qualitative data was created and retrieved and to clearly describe the nature and meaning of the units of observation and metadata elements available as a part of the dataset. The file should also contain a detailed protocol used to collect and organize data or a link to the source code of the script used for data collection and organization. All this will aid researchers in understanding better the nature of the dataset and interpreting findings derived from the dataset by relating it back to the context within which the data was created and retrieved. This information will also contribute to the reproducibility and extendibility of the findings derived from this data as other researchers would be in a better position to understand and replicate the exact steps used for data collection.

4. The Case of Dice.com

In a similar approach to Krotov et al. (2023), data obtained from Dice.com is used to illustrate making qualitative data tidy. Dice.com (www.dice.com) is a job posting site for IT and engineering professionals (DHI Group, 2023). The purpose of this illustrative, qualitative research project is to answer the following research question: “What are the most in-demand skills for the role of a systems analyst?” Data collection is performed using the R programming language as a technology platform. The language has grown in popularity in both research and industry data science projects, and numerous packages for data collection, processing, and analysis are available in R. Among them are R packages developed specifically for web data collection.

The purpose of this illustrative, qualitative research project is to answer the following research question: "What are the most in-demand skills for the role of a systems analyst?" It may be possible to answer this question based on data available on Dice.com.

4.1 Phase 1: Analyzing the Website

There are several steps involved in the analysis of the Dice website. In the first step, social structures used to create data are analyzed so that one can understand what impact they can have on data quality. Second, the API manual provided by the website is analyzed for the programmatic retrieval of the website's data. As a third step, the structure of web pages that contain complete job descriptions is analyzed to determine which page elements contain qualitative job descriptions. This step results in a better understanding of how data can be retrieved programmatically to answer the research question.

Dice is a leading, specialized recruitment website (Lee, 2007). The website specializes in job search and recruitment solutions for IT and engineering workers (DHI Group, 2023). The website contains close to 100,000 active job listings available to the public, as well as 2.1 million résumés that are not accessible to the public. Each month, the website reports more than 2 million unique visitors. A detailed examination of sample job listings on Dice reveals that the data is of high quality. It's not surprising since Dice specializes in IT recruitment data products. This means that the data on the website is professionally managed. Furthermore, job listings are created by professionals, such as managers and recruiters.

Using Dice's API, we are now able to determine what data is available and how it is organized and managed. The API output does not directly provide job descriptions. The API, however, returns URLs for accessing detailed job descriptions. To determine which HTML element contains the job descriptions, Google Chrome's Inspect Element feature and SelectorGadget add-on are used. This information is needed to develop an R script that automates Dice data collection.

4.2 Phase 2: Retrieving Raw Data

There are three main steps to retrieving job listings related to systems analyst from Dice. The first step is to finalize the links between units of observation and units of analysis. As part of the second step, an R script is developed and debugged. This script “crawls” the website and downloads data related to systems analyst job listings based on the study's units of analysis (or theoretical concepts). Before the script was used to retrieve data, it was "pretested" several times on smaller tasks (e.g., niche jobs in areas known to contain very few job listings). The third step involves running the script with some degree of human supervision to retrieve the data and save it into a data frame object in R.

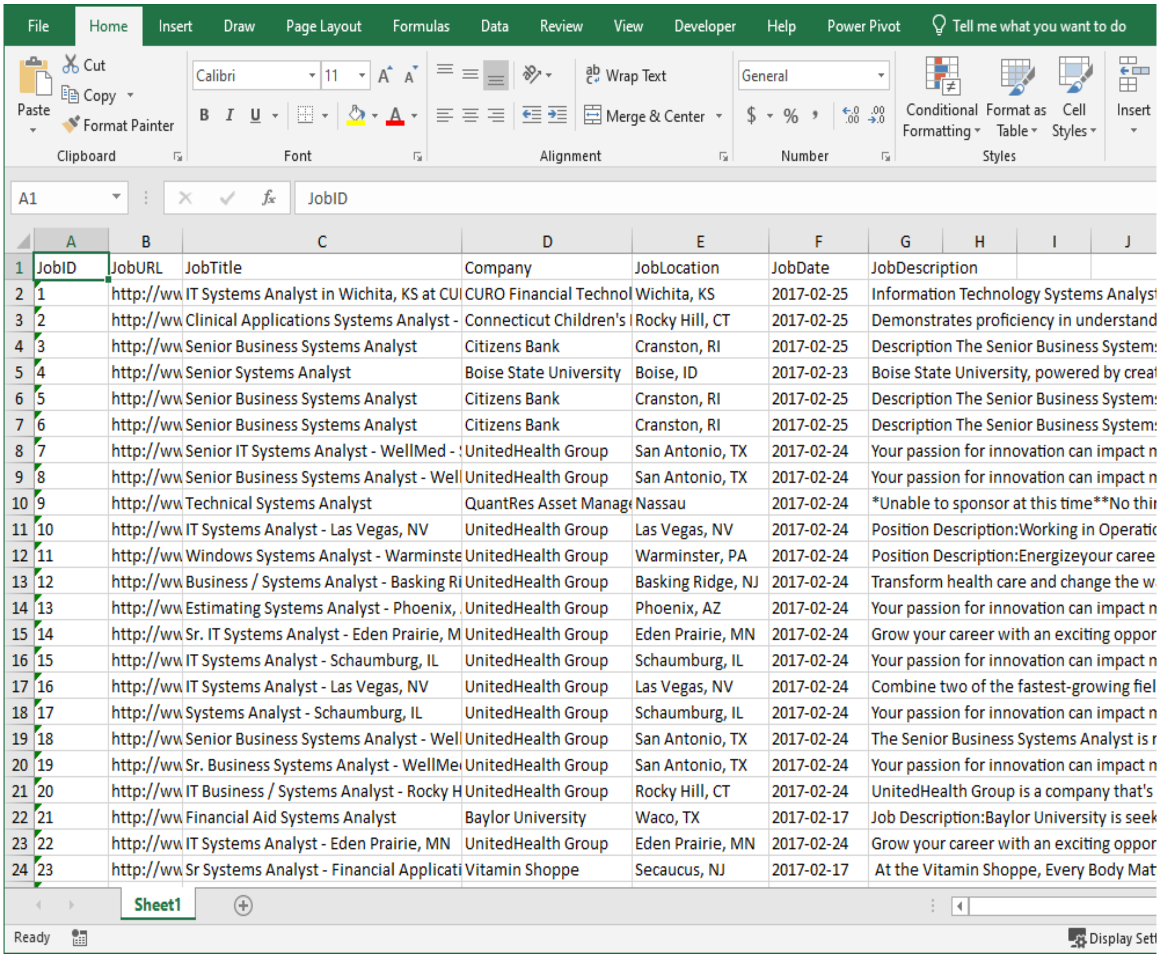
A data frame is probably the most common type of data structure in R. A data frame corresponds roughly to an Excel table, where (a) each column represents a variable; (b) each variable has a name (or column heading); (c) each column can be of a single data type; and (d) each row represents an observation. This data frame is first saved to the R project environment (i.e., in Random Access Memory). This data can be saved and distributed along with the project file and analyzed within the same project. However, the goal of this illustrative example is to save the data in tidy qualitative data format in an Excel file, so that it can be analyzed using other software tools (e.g., NVivo). The next section describes how the data is “tidied up” further and saved in an Excel file containing some metadata.

4.3 Phase 3: Tidying Raw Data

The following steps are involved in tidying raw data. In the first step, meaningful linguistic labels are created for the columns in the data frame (Criterion 6). The second step is to save the data in a tidy Excel file to allow subsequent analysis by other software tools or even by hand (Criterion 1). For researchers to track scraping jobs, a date and time stamp is added to the file name to provide metadata about the scraping operation (Criterion 5). In the third step, README.rtf is added to describe the dataset and its elements (Criterion).

Step 1: Creating Linguistic Labels

Before saving the job description data into files, the following column labels were created: JobID, JobURL, JobTitle, Company, JobLocation, JobDate, and JobDescription (see Figure 3). These labels are short yet contain enough information for someone to grasp what these fields store. Each row of the data depicted in Figure 3 represents one unit of observation, which is a job description posted to Dice.com (Criterion 3). Since each Excel sheet is formatted as a table, each unit of observation is clearly demarcated from the rest (Criterion 4).



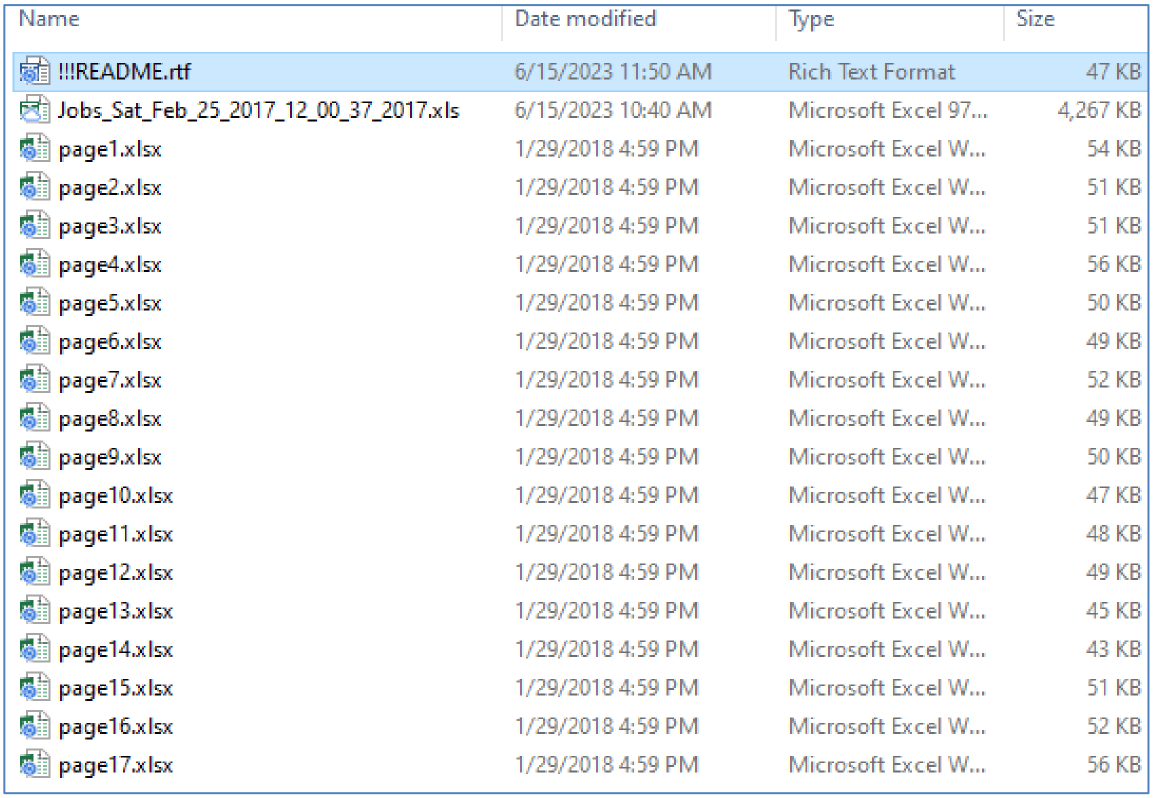
A1 JobID

	A	B	C	D	E	F	G
1	JobID	JobURL	JobTitle	Company	JobLocation	JobDate	JobDescription
2	1	http://ww	IT Systems Analyst in Wichita, KS at CU	CURO Financial Technol	Wichita, KS	2017-02-25	Information Technology Systems Analyst
3	2	http://ww	Clinical Applications Systems Analyst -	Connecticut Children's	Rocky Hill, CT	2017-02-25	Demonstrates proficiency in understand
4	3	http://ww	Senior Business Systems Analyst	Citizens Bank	Cranston, RI	2017-02-25	Description The Senior Business System
5	4	http://ww	Senior Systems Analyst	Boise State University	Boise, ID	2017-02-23	Boise State University, powered by creat
6	5	http://ww	Senior Business Systems Analyst	Citizens Bank	Cranston, RI	2017-02-25	Description The Senior Business System
7	6	http://ww	Senior Business Systems Analyst	Citizens Bank	Cranston, RI	2017-02-25	Description The Senior Business System
8	7	http://ww	Senior IT Systems Analyst - WellMed -	UnitedHealth Group	San Antonio, TX	2017-02-24	Your passion for innovation can impact n
9	8	http://ww	Senior Business Systems Analyst - Well	UnitedHealth Group	San Antonio, TX	2017-02-24	Your passion for innovation can impact n
10	9	http://ww	Technical Systems Analyst	QuantRes Asset Manag	Nassau	2017-02-24	*Unable to sponsor at this time**No thi
11	10	http://ww	IT Systems Analyst - Las Vegas, NV	UnitedHealth Group	Las Vegas, NV	2017-02-24	Position Description:Working in Operati
12	11	http://ww	Windows Systems Analyst - Warminste	UnitedHealth Group	Warminster, PA	2017-02-24	Position Description:Energizeyour caree
13	12	http://ww	Business / Systems Analyst - Basking Ri	UnitedHealth Group	Basking Ridge, NJ	2017-02-24	Transform health care and change the w
14	13	http://ww	Estimating Systems Analyst - Phoenix,	UnitedHealth Group	Phoenix, AZ	2017-02-24	Your passion for innovation can impact n
15	14	http://ww	Sr. IT Systems Analyst - Eden Prairie, M	UnitedHealth Group	Eden Prairie, MN	2017-02-24	Grow your career with an exciting oppor
16	15	http://ww	IT Systems Analyst - Schaumburg, IL	UnitedHealth Group	Schaumburg, IL	2017-02-24	Your passion for innovation can impact n
17	16	http://ww	IT Systems Analyst - Las Vegas, NV	UnitedHealth Group	Las Vegas, NV	2017-02-24	Combine two of the fastest-growing fiel
18	17	http://ww	Systems Analyst - Schaumburg, IL	UnitedHealth Group	Schaumburg, IL	2017-02-24	Your passion for innovation can impact n
19	18	http://ww	Senior Business Systems Analyst - Well	UnitedHealth Group	San Antonio, TX	2017-02-24	The Senior Business Systems Analyst is r
20	19	http://ww	Sr. Business Systems Analyst - WellMe	UnitedHealth Group	San Antonio, TX	2017-02-24	Your passion for innovation can impact n
21	20	http://ww	IT Business / Systems Analyst - Rocky H	UnitedHealth Group	Rocky Hill, CT	2017-02-24	UnitedHealth Group is a company that's
22	21	http://ww	Financial Aid Systems Analyst	Baylor University	Waco, TX	2017-02-17	Job Description:Baylor University is seek
23	22	http://ww	IT Systems Analyst - Eden Prairie, MN	UnitedHealth Group	Eden Prairie, MN	2017-02-24	Grow your career with an exciting oppor
24	23	http://ww	Sr Systems Analyst - Financial Applicati	Vitamin Shoppe	Secaucus, NJ	2017-02-17	At the Vitamin Shoppe, Every Body Mat

Figure 3. The Content of the Main File

Step 2: Saving Data in Excel Files

This step involves using R to produce the file structure depicted in Figure 4. As one can see in Figure 4, separate Excel files were created for each JSON-formatted page of job listings scraped. By default, the API returns information regarding 50 job listings at a time. Therefore, each of the 50 job listings is also stored in its own separate Excel file. The saved pages collectively duplicate the information stored in the main Excel sheet (Jobs_Sat_Feb_25_12_00_37_2017.xlsx). Duplication is done to prevent data loss. Datasets with thousands of observations may take a computer hours or even days to collect. In the event that script execution is halted (e.g., the web server fails, the computer runs out of memory or freezes, the building goes dark) the researchers still have access to the data collected before the failure. Until the script execution is halted, these data will be saved as Excel files. The main, time-stamped Excel file (Jobs_Sat_Feb_25_12_00_37_2017.xlsx) shown in Figure 4 contains the entire dataset. The entire dataset is saved only if the entire web scraping job is completed. As indicated by the file name, the data were saved on Feb. 25, 2017, at 12:00:37 p.m. As one can see in Figure 3, the data saved in this file is “clean”: It’s free from such “impurities” as extra white spaces, unreadable characters, markup tags, and repeated or redundant elements. (Criterion 1).



Name	Date modified	Type	Size
!!!README.rtf	6/15/2023 11:50 AM	Rich Text Format	47 KB
Jobs_Sat_Feb_25_2017_12_00_37_2017.xls	6/15/2023 10:40 AM	Microsoft Excel 97...	4,267 KB
page1.xlsx	1/29/2018 4:59 PM	Microsoft Excel W...	54 KB
page2.xlsx	1/29/2018 4:59 PM	Microsoft Excel W...	51 KB
page3.xlsx	1/29/2018 4:59 PM	Microsoft Excel W...	51 KB
page4.xlsx	1/29/2018 4:59 PM	Microsoft Excel W...	56 KB
page5.xlsx	1/29/2018 4:59 PM	Microsoft Excel W...	50 KB
page6.xlsx	1/29/2018 4:59 PM	Microsoft Excel W...	49 KB
page7.xlsx	1/29/2018 4:59 PM	Microsoft Excel W...	52 KB
page8.xlsx	1/29/2018 4:59 PM	Microsoft Excel W...	49 KB
page9.xlsx	1/29/2018 4:59 PM	Microsoft Excel W...	50 KB
page10.xlsx	1/29/2018 4:59 PM	Microsoft Excel W...	47 KB
page11.xlsx	1/29/2018 4:59 PM	Microsoft Excel W...	48 KB
page12.xlsx	1/29/2018 4:59 PM	Microsoft Excel W...	49 KB
page13.xlsx	1/29/2018 4:59 PM	Microsoft Excel W...	45 KB
page14.xlsx	1/29/2018 4:59 PM	Microsoft Excel W...	43 KB
page15.xlsx	1/29/2018 4:59 PM	Microsoft Excel W...	51 KB
page16.xlsx	1/29/2018 4:59 PM	Microsoft Excel W...	52 KB
page17.xlsx	1/29/2018 4:59 PM	Microsoft Excel W...	56 KB

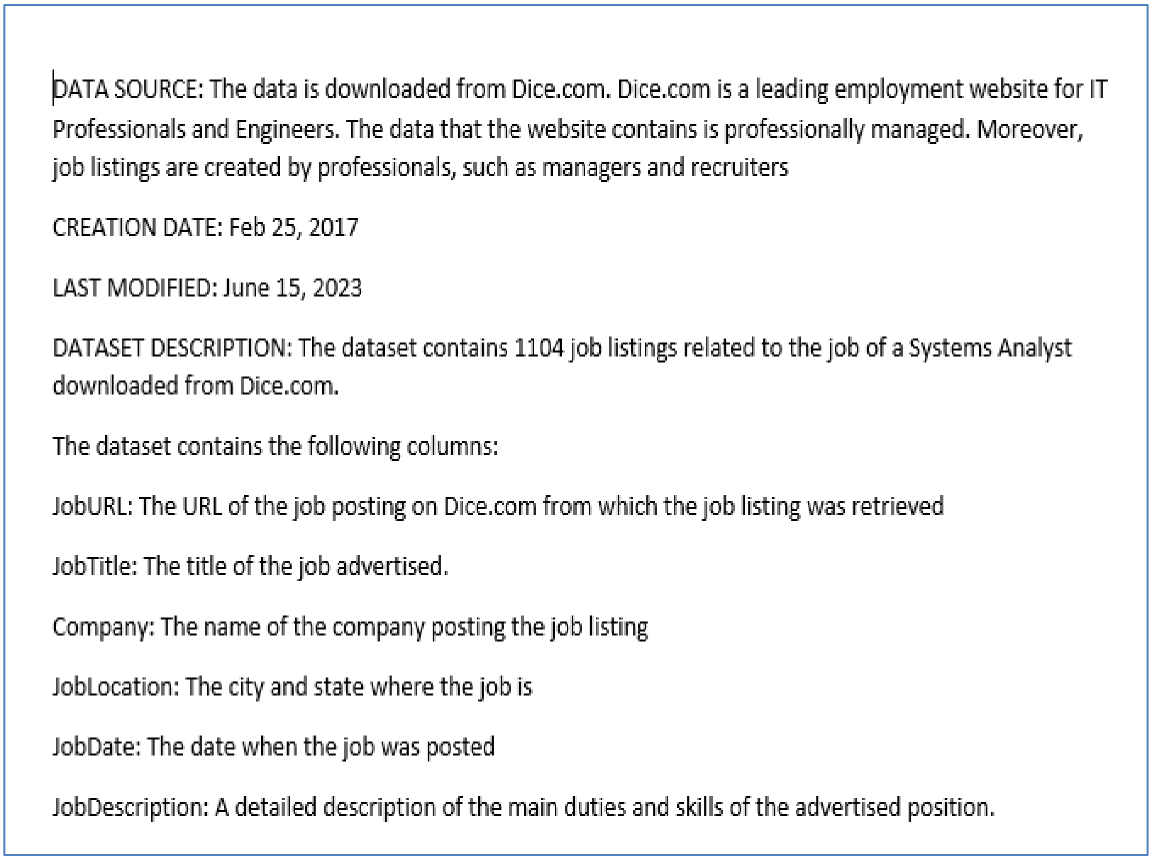
Figure 4. The File Structure Generated

Step 3: Creating a "README" File

Finally, a dataset description file named !!!README.rtf is added (Criterion 7). This file is added to the file structure depicted in Figure 4 to capture the context within which the data was collected and provide brief descriptions of the data and metadata elements comprising the dataset. The content of this !!!README.rtf file is provided in Figure 5.

4.4. Reflecting on Data "Tidiness" of the Dataset

Note that the data stored in the main file (Jobs_Sat_Feb_25_12_00_37_2017.xlsx) is "tidy" (see Figure 3), since it meets all seven tidy data criteria discussed earlier. First, the file is saved in Excel format, making it easy to access and analyze the data manually and programmatically. Second, the dataset is clean: It is free from extra white spaces, HTML tags, and other "impurities" that often accompany a dataset retrieved from the web. Third, the data is organized in accordance with the unit of observation, which is a job listing for a systems analyst position. Fourth, each unit of observation is clearly delineated: Each row corresponds to a single job listing. Fifth, each unit of observation, which is a job listing, is supplemented with relevant metadata elements in the form of columns. These metadata elements include JobID (which is a primary key for each of the observations), JobURL (the URL used to obtain a job description), JobTitle (the job title of the position listed), Company (the name of the company that posted the job), JobDate (the date when the job was posted), and JobDescription (the full description of the job advertised) (see Figure 3). These metadata elements supplement each unit of analysis and enrich the understanding of the context within which the data was collected and should be interpreted. Sixth, the labels used for the metadata elements are short, simple, and easy to understand. Finally, a !!!README.rtf file is added to the file structure to provide a summary of the dataset together with its main complements (see Figure 5).



DATA SOURCE: The data is downloaded from Dice.com. Dice.com is a leading employment website for IT Professionals and Engineers. The data that the website contains is professionally managed. Moreover, job listings are created by professionals, such as managers and recruiters

CREATION DATE: Feb 25, 2017

LAST MODIFIED: June 15, 2023

DATASET DESCRIPTION: The dataset contains 1104 job listings related to the job of a Systems Analyst downloaded from Dice.com.

The dataset contains the following columns:

JobURL: The URL of the job posting on Dice.com from which the job listing was retrieved

JobTitle: The title of the job advertised.

Company: The name of the company posting the job listing

JobLocation: The city and state where the job is

JobDate: The date when the job was posted

JobDescription: A detailed description of the main duties and skills of the advertised position.

Figure 5. The Content of the !!!README File

5. Directions for Future Research

The focus of the tidy qualitative data and criteria is on data organization, that is, the format in which qualitative data retrieved from the web data should be organized and stored. The seven criteria of tidy qualitative data, if met, do not guarantee that the content or meaning of the collected data will be adequate for answering a research question (or questions). Specifically, there are two important content-related considerations that need to be explored in relation to any dataset collected from the web: data privacy and data validity (Krotov & Johnson, 2023).

Web data often contains personally identifiable information (PII) that can potentially compromise the privacy of individuals and reveal important trade secrets of organizations affiliated with a particular dataset (Krotov & Johnson, 2023). Even if web data is seemingly anonymous, it can still reveal confidential information if this data is triangulated via other data sources (Ives & Krotov, 2006). Further research is needed to develop guidelines on how to anonymize and mask data, so that (a) the privacy of the stakeholders affiliated with a dataset is preserved, and (b) important contextual information is not lost as a result of data masking and anonymization.

Another important area linked to the concept of tidy qualitative data not addressed in this paper is the degree to which digital text downloaded from the web carries valid information. The problem is that virtually all modern forms of digital communication, such as emails, text messages, social media posts, product reviews, and online news articles, often contain wrong or misleading information—either intentionally or unintentionally (George & Luo, 2019). The recent explosion in the availability of GAI tools makes the creation of fake or low quality digital content easier than ever (Bansal et al, 2023). For example, a dataset containing "fake" online product reviews can be "tidied up" in accordance with the principles discussed in this article but still be of questionable value in social sciences research. Reflection on the social structures used to generate data can give researchers some reassurance in relation to its reliability and accuracy (Krotov et al., 2020). Still, the question of the extent to which this data is usable in academic research should be explored and confirmed through more reliable means, and this can be another area to explore in relation to the tidy data concept.

While data privacy and data validity are very important considerations for qualitative researchers, both issues are somewhat detached from the main focus of this paper, which is data organization. Data privacy and data validity are, rather, related to the content or meaning of the data retrieved. Thus, while bringing these issues to the readers' attention, this paper does not attempt to address issues related to these two important considerations. A thorough exploration of these issues is left for further research.

6. Conclusion

The qualitative data available on the web presents social scientists with a variety of opportunities for answering important research questions with rigor, precision, and timeliness. This data can be retrieved using a variety of manual, semiautomated, and fully automated tools and approaches. While every web data sourcing project is unique and often requires customized web scraping tools, all approaches to web data sourcing should result in a tidy qualitative dataset. The main contribution of this paper is providing guidance to researchers on how to make a qualitative dataset tidy. This paper proposes the following definition of tidy qualitative data: qualitative data that is tidy when it is organized and stored in a way that preserves its context and makes it easy to analyze with a variety of qualitive techniques. This definition is operationalized by proposing that a tidy qualitative dataset should meet the following seven criteria: 1. the data needs to be stored in a format easily accessible by humans and computers; 2. the data should be clean; 3. the data should be organized according to the unit of observation; 4. each observation should be clearly demarcated from other observations; 5. metadata elements should be added to each unit of observation; 6. the data and metadata elements should be labeled properly; 7. the dataset should be supplemented with a description file. A tidy qualitative dataset compliant with the specific criteria discussed in this paper allows for a wide variety of analytical approaches being used to analyze the dataset. This, in turn, will help researchers answer a variety of old and new research questions with more rigor and precision, taking full advantage of the nuggets of gold in the oceans of qualitative data floating around the web.

7. References

Bansal, G., Hosack, B., Iversen, J., Mitchell, A., Hadidi, R., & George, J. (2023). ChatGPT–another hype or out-of-this-world? *Journal of the Midwest Association for Information Systems (JMWAIS), 2023*(2), Article 3, 29-35.

Cisco Systems. (2016). *Cisco visual networking index: Forecast and methodology* [White paper]. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.html

Codd, E. F. (1972). Further normalization of the data base relational model. *Data Base Systems, 6*, 33-64.

DHI Group. (2023). *About Us.* https://dhigroupinc.com/about-dhi/default.aspx

Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review*, *14* (4), 532-550.

George, J. F., & Luo, J. (2019). The effects of media differences and expertise on deception detection accuracy. *Journal of the Midwest Association for Information Systems (JMWAIS)*, *2019*(1), Article 5, 69-80.

Goes, P. B. (2014). Editor's comments: Big data and IS research. *MIS Quarterly*, *38*(3), iii-viii.

Hadidi, R., & Klein, B. D. (2024). The future of work, physical location of workers, technological issues and implications. *Journal of the Midwest Association for Information Systems (JMWAIS), 2024*(1), Article 1, 1-7.

Ives, B., & Krotov, V. (2006). Anything you search can be used against you in a court of law: Data mining in search archives. *Communications of the Association for Information Systems*, *18*(1), 29.

Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis,* (33), 171-185.

Klein, H. K., & Myers, M. D. (1999). A set of principles for conducting and evaluating interpretive field studies in information systems. *MIS Quarterly, 23*(1), 67-94.

Krotov, V., & Johnson, L. (2023). Big web data: Challenges related to data, technology, legality, and ethics. *Business Horizons*, *66*(4), 481-491.

Krotov, V., Johnson, L., & Silva, L. (2020). Tutorial: Legality and ethics of web scraping. *Communications of the Association for Information Systems*, 47, 539-563.

Krotov, V. & Silva, L. (2018). Legality and ethics of web scraping. *Twenty-Fourth Americas Conference on Information Systems.*

Krotov, V., & Tennyson, M. (2018). Scraping financial data from the web using R language. *Journal of Emerging Technologies in Accounting, 15*(1), 169-181.

Lee, I. (2007). An architecture for a next-generation holistic e-recruiting system. *Communications of the ACM, 50*(7), 81-85.

Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. (2015). *Automated data collection with R: A practical guide to web scraping and text mining*. John Wiley & Sons, Ltd.

Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226-1227.

Schutt, R., & O'Neil, C. (2013). *Doing data science: Straight talk from the frontline*. O'Reilly Media, Inc.

Triche, J., & Walden, E. (2018). The use of impression management strategies to manage stock market reactions to IT failures. *Journal of the Association for Information Systems, 19*(4), 333-357.

Vaast, E., Safadi, H., Lapointe, L., & Negoita, B. (2017). Social media affordance for connective action: An examination of microblogging use during the Gulf of Mexico oil spill. *MIS Quarterly, 41*(4), 1179-1206.

Walsham, G. (1995). Interpretive case studies in IS research: Nature and method. *European Journal of Information Systems, 4*(2), 74-81.

Watson, H. J. (2014). Tutorial: Big data analytics: Concepts, technologies, and applications. *Communications of the Association for Information Systems, 34*(1), 1247-1268.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1-23.

Yin, R. K. (1994). *Case study research: Design and methods*, (2nd ed., Vol. 5). SAGE Publications.

Author Biography

Vlad Krotov received his PhD in Management Information Systems from the Department of Decision and Information Sciences, University of Houston (USA). Currently, he is a Professor of Information Systems at Murray State University and a consultant at Accreditation.Biz - an international accreditation consulting company for business schools. He is also a founder of ScrumEducator.org – a free resource for training students, educators, and professionals in essential soft skills. His research and teaching interests are at the intersection of business and technology and fall under such topics as business analytics, strategic IT management, project management, innovation, and ethical aspects of AI.