Journal of International Technology and Information Management

Volume 27 | Issue 3 Article 7

1-1-2019

Inside the Black Box of Dictionary Building for Text Analytics: A Design Science Approach

Qi Deng
Carleton University, qideng3@cmail.carleton.ca

Michael J. Hine
Carleton University, mike.hine@carleton.ca

Shaobo Ji
Carleton University, shaobo.ji@carleton.ca

Sujit Sur
Sprott School of Business, Carleton University, sujit.sur@carleton.ca

Inside the Black Box of Dictionary Building for Text Analytics: A Design Science Approach

Qi Deng
Michael J. Hine
Shaobo Ji
Carleton University

Sujit Sur
Sprott School of Business, Carleton University

ABSTRACT

The purpose of this paper is to develop and demonstrate a dictionary building process model for text analytics projects following the design science methodology. Using inductive consensus-building, we examined prior research to develop an initial process model. The model is subsequently demonstrated and validated by using data to develop an environmental sustainability dictionary for the IT industry. To our knowledge, this is an initial attempt to provide a normalized dictionary building process for text analytics projects. The resulting process model can provide a road map for researchers who want to use automated approaches to text analysis but are currently prevented by the lack of applicable domain dictionaries. Having a normalized design process model will assist researchers by legitimizing their work requiring dictionary building and help academic reviewers by providing an evaluation framework. The resulting environmental sustainability dictionary for IT industry can be used as a starting point for future research on Green IT and sustainability management.

KEYWORDS: dictionary building, process, environmental sustainability, text analytics, design science, IT industry

ISSN: 1941-6679-On-line Copy

INTRODUCTION

Text analytics provides an efficient method to understand unstructured text, allowing researchers to systematically assess different aspects of the core concept(s) they are interested in. Many text analytic projects are reliant on thesaurus-like dictionaries, which consist of categories that contain lists of entries (i.e., words, word stems, or phrases) with shared meanings (Landmann & Zuell, 2008; Weber, 1983). For example, in Stone, Dumphy, and Ogilvie (1966) psycho-sociological dictionary, the concept/category of *self* is described by the words, *I*, *me*, *my*, *mine*, and *myself*, and the concept/category of *selves* by the words, *we*, *us*, *our*, *ours*, and *ourselves*. To analyze a corpus, the frequencies of the entries and categories are counted and, based on these frequencies, the relative importance or changes over time of the central concepts in the text can be determined. Text analytics is being increasingly embraced by researchers because of its ability to process large volumes of data at high speed (Krippendorff, 2004). Such ability is particularly important in the current context of big data. Compared with manual content analysis, the text analytics approach is "*consistent (without random human error), replicable (the process is rule-based), scalable (coding efforts are the same regardless of the number of reports analyzed), and transparent (when the keywords/phrases and search criteria used to automate identification are made available)*" (Boritz, Hayes & Lim, 2013).

In dictionary-based text analytic projects, the quality of the results is dependent on the quality of the dictionary (Laver & Garry, 2000). Thus, a main challenge for researchers is to develop a satisfactory dictionary (Wiedemann, 2013). Developing a special-purpose dictionary is a formidable, iterative, and time-consuming process which could last from months to years (Brier & Hopp, 2011; Landmann & Zuell, 2008; Morris, 1994; Péladeau & Stovall, 2005; Schrodt & Gerner, 2012). Because of this, researchers and practitioners rely on available dictionaries, rather than build their own (Krippendorff, 2004). Unfortunately, generic dictionaries often provide little insight into the underlying thematic structure of a domain specific corpus of documents. Additionally, given the changing meanings of words over time and space, existing dictionaries might need to be adapted before being applied. Therefore, developing a dictionary for one's own research purposes is often necessary. Once well-developed, a dictionary can be applied to any text in a similar domain with little additional effort, and thus, a number of content analyses would benefit from this (Boritz et al., 2013; Brier & Hopp, 2011; Péladeau & Stovall, 2005). Given the importance of dictionary building, it is surprising that the process of developing a dictionary has not received proper attention. Although Laver and Garry (2000, p. 626) indicated that, "*what remains constant over time is thus the dictionary generation procedure, not the actual word lists in the dictionary*", to our

knowledge, no research has tried to standardize the dictionary building process. The lack of a standardized process contributes to dictionary building being criticized for its 'abductive manner' (Wiedemann, 2013).

The aim of this study is to develop, evaluate and demonstrate a process model for dictionary building to be used in text analytics projects. The contributions of this paper are threefold. First, this paper is an initial attempt at providing a standardized dictionary building process. Second, the dictionary building process proposed in this paper helps provide a road map for researchers who want to use text analytics but are constrained by the lack of available dictionaries. It also helps researchers by legitimizing their research that is on, or dependent on, dictionary building and assists academic reviewers by providing an evaluation framework. Third, the standardized process could promote research on dictionary building and on research that is reliant on building a dictionary and thus facilitate the proliferation of the text analytics method.

This paper is organized following the design science research publication schema proposed by Gregor & Hevner (2013). Section two presents the prior work on dictionary building. Section three presents the method employed to develop the dictionary building process. Section four provides a concise description of the artifact, which in this case, is the dictionary building process model. Section five evaluates the usefulness of the artifact through multiple forms of validation and demonstration on how the process model can be used to develop an environmental sustainability dictionary for IT companies. Section six provides a discussion on the dictionary building process. Section seven presents the conclusions of this study.

LITERATURE REVIEW

This section contains three subsections including: 1) a review of the dictionaries built in prior research; 2) a framework that details different approaches to dictionary building, including a comparison of their relative advantages and disadvantages; and 3) a re-positioning of the dictionary building process through the lens of design science research. The purpose of the three subsections are: 1) to understand what has been done in previous research, 2) to establish an appropriate scope for this study, and 3) to provide a theoretical foundation for developing the dictionary building process model.

ISSN: 1941-6679-On-line Copy

A Review of Existing Dictionaries

To build a dictionary, one needs to manually or automatically identify the 'right' words and/or phrases in the corpus and assign them into different categories that represent concepts that the researcher is interested in. For example, to build a sentiment dictionary which can be used to analyze online product reviews, researchers may identify the words "satisfy", "good", and "useful" as being representative of positive sentiment and the words "terrible", "angry", and "useless" as that of negative sentiment. Since the 1960s, researchers have been developing dictionaries for various purposes (Loughran & McDonald, 2011; Schwartz & Ungar, 2015; Young & Soroka, 2012). Now, numerous dictionaries, varying widely with respect to languages, categories, and scope of coverage have been used for research (see Table 1).

Table 1. Dictionaries built in previous studies

Source	Dictionary Domain	Dictionary Structure
Aaldering & Vliegenthart (2015)	Dictionary (Dutch) of public leadership image in newspapers	Not specified
Abrahamson & Eisenman (2008)	Dictionary of rational and normative words in the language of employee-management techniques	1781 entries/2 categories (23 sub-categories)
Albaugh, Sevenans, Soroka, & Loewen (2013)	Dictionary (English and Dutch) of policy agendas	Not specified
Bengston & Xu (1995)	Dictionary of forest values	4 categories
Boritz et al. (2013)	Dictionary of IT context indicator and dictionary of IT weaknesses	1 category/14 categories
Cohen (2012)	Dictionary of cognitive rigidity	250 entries/2 categories
Debortoli, Müller, & vom Brocke (2014)	Dictionary of competency-related terms in business intelligence and big data job ads	1570 entries
de-Miguel-Molina, Chirivella-González, & García-Ortega (2016)	Dictionary of corporate philanthropy	6 categories

Guo, Vargo, Pan, Ding, & Ishwar (2016)	Dictionary of news topics and public opinions of U.S. political elections	16 categories
Hart (1984, 2000)	DICTION: four major dictionaries and seven minor dictionaries.	Not specified
Hiller, Marcotte, & Martin (1969)	Dictionary of characteristics of writing style	280 entries/3 categories
Kirilenko, Stepchenkova, Romsdahl, & Mattis (2012)	Dictionary of precautionary principle	Not specified
König & Finke (2013)	Dictionary (German) of counterterrorist content	57 words/1 category
	Dictionary (German) of partisan security and civil liberties preferences	1678 words/2 categories
Laver & Garry (2000)	Dictionary of policy position.	Not specified
Lesage & Wechtler (2012)	Dictionary of auditing research topics	481 entries
Loughran & McDonald (2011)	Dictionary of tone in financial text	3752 entries/6 categories
Martindale (1975, 1990)	Regressive imagery dictionary	5336 words/68 categories
Matthies & Coners (2015)	Dictionary of corporate risks	89/6 categories
Mergenthaler (1996, 2008)	Dictionary of emotion tone; Dictionary of abstraction.	2305 entries/4 categories; 3900 entries
Opoku, Abratt, & Pitt (2006)	Dictionary of business school brand personality	1625 words/5 categories
Park, Lu & Marion (2009)	Dictionary of job description	3 categories
Pennebaker, Boyd, Jordan, & Blackburn (2015)	Linguistic Inquiry and Word Count (LIWC 2015)	6,400 words, word stems, and select emoticons
Péladeau & Stovall (2005)	Dictionary of aviation safety	Not specified
Rooduijn & Pauwels (2011)	Dictionary of anti-elitism.	75 entries/8 categories

ISSN: 1941-6679-On-line Copy

Smith & Chang (1996)	Dictionary of online image and video subject	Not specified
Strapparave & Valitutti (2004)	WordNet-Affect: dictionary of affective concepts	2874 synsets and 4787 words
Vasalou, Gill, Mazanderani, Papoutsi, & Joinson (2011); Gill, Vasalou, Papoutsi, & Joinson (2011)	Dictionary of privacy related issues	355 entries/8 categories
Wade, Porac, & Pollock (1997)	Dictionary of compensation justification	94 entries/5 categories
Whissell (1986)	Dictionary of affect in language	4323 words
Wilson (2006)	Dictionary of body type.	778 entries/2 categories
Young & Soroka (2012)	Dictionary of sentiment in political communication.	4567 entries/2 categories

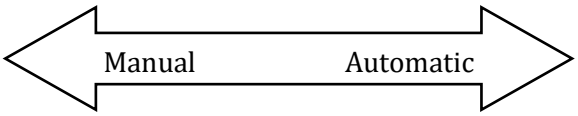
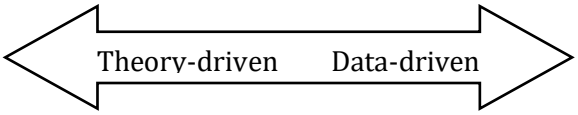
Most dictionaries are generated for a particular purpose or genre of text, and as a consequence tend to be temporally and corporally specific (Young & Soroka, 2012). Thus, developing new dictionaries, or, at least, adapting existing dictionaries, is unavoidable.

Approaches to Dictionary Building: A Spectrum from Manual to Automatic

The majority of previous research using a dictionary has paid scant attention to the processes followed in developing the dictionaries themselves. In general, the dictionary building processes are described in a very perfunctory way and no systematic and normalized dictionary building process has been proposed. However, there is some dictionary building guidance that can be culled from a thorough overview of previous research and some general discussions on dictionary building has been provided (see Brier & Hopp, 2011; Cohen, 2012; Krippendorff, 2004; Schwartz & Ungar, 2015; Young & Soroka, 2012). Through summarizing these discussions in previous research (i.e., Brier & Hopp, 2011; Cohen, 2012; Krippendorff, 2004; Schwartz & Ungar, 2015; Young & Soroka, 2012), we developed a framework to distinguish between three different dictionary building approaches and their characteristics (See Table 2).

ISSN: 1941-6679-On-line Copy

Table 2. Three dictionary building approaches: a comparison

		Manual	Semi-Automatic	Automatic
Activity	• *Developing categories* • *Identifying entries* • *Categorizing entries*			
Approach	• *Direction*			
Requirement	• *Domain knowledge*	High	Moderate	Low
	• *Programing knowledge*	Low	Moderate	High
Capability	• *Corpus size*	Low	Moderate	High
	• *Dictionary size*	Low	Moderate	High
Outcome	• *Dictionary abstraction*	High	Moderate	Low
	• *Dictionary variation*	Low	Moderate	High

Three approaches to dictionary building have been identified: 1) manual; 2) semi-automatic; and 3) automatic. We distinguish between them by the automaticity of the three core activities in the dictionary building process: 1) developing categories; 2) identifying entries; 3) categorizing entries. Each of the three activities could be manual, semi-automatic, or automatic. If all three activities of a dictionary building process were purely manual (automatic), then the process would be viewed as the manual (automatic) approach; otherwise, the process would be viewed as semi-automatic.

Rooted in the traditional content analysis method, manual dictionary building is usually a theory-driven process, which is similar to the process of developing a coding schema. Since the core activities are conducted manually, this approach requires the highest domain knowledge and the lowest programing knowledge. In addition, it does not rely on a large corpus, and typically results in dictionaries with small sizes. Because it is a theory-driven process, the manual dictionary building

ISSN: 1941-6679-On-line Copy

approach usually results in dictionaries with high abstractions and low variations. The dictionaries developed using a manual approach usually have a theory-based and systematic category structure and are less probable to have unexpected categories or entries.

Automatic dictionary building is rooted in the field of computational linguistics which focuses on modeling language (Jurafsky & Martin, 2000) and has mainly been applied in the field of Medical Science and Bioinformatics. In general, it involves extracting key words and/or phrases automatically based on learning algorithms and subsequently evaluating the resulting dictionary through experiments or comparing it with existing dictionaries. Social sciences have just recently begun to adopt this method because of previous challenges associated with the large sample size requirement and its low methodological accessibility (Schwartz & Ungar, 2015). Compared to the manual approach, automatic dictionary building requires the lowest domain knowledge, but the highest programing knowledge. It can handle very large corpora and produce 'big' dictionaries. However, because it is a data-driven process, the resulting dictionaries may not correspond to theory and can result in unexpected categories and/or entries.

In the semi-automatic approach, researchers conduct the three activities and make their own judgments with the assistance of text analysis software. For example, to develop a category structure, researchers could initially propose or adopt some categories based on theory and then modify them based on the result of automatic topic extraction from a corpus. To identify the entries, one can first narrow down the scope of the corpus by setting up a frequency criterion with the help of text analysis software.

Each of the three dictionary building approaches has its advantages and disadvantages, and the choice of the appropriate one should be made based on the objectives of the research project under consideration. In this paper, we focus on semi-automatic dictionary building for three reasons. First, it is the most widely-adopted dictionary building approach (Brier & Hopp, 2011; Schwartz & Ungar, 2015). Second, the semi-automatic approach can potentially leverage existing theoretical bases and the contents of the corpus itself in executing the three dictionary building activities. Third, although the semi-automatic approach is not as computationally efficient as the automatic approach, it is self-justified by its accessibility: one does not need a programing background to adopt it. Therefore, our objective in this paper is to develop a process model for semi-automatic dictionary building.

ISSN: 1941-6679-On-line Copy

Rethinking the Dictionary Building via the Lens of Design Science Research

Design science is a research paradigm that focuses on problem-solving (March & Storey, 2008). It aims to create artifacts (i.e., construct, model, method, or instantiation) to solve identified problems and serve human purposes (Hevner et al., 2004; March & Smith, 1995; March & Storey, 2008; Simon, 1996). According to March & Smith (1995), the core activities of design science research are 'build' (construct an artifact for a specific purpose) and 'evaluate' (determine how well the artifact performs). The dictionary building process can be framed as a design problem and thus can be addressed by the design science research method. Through this lens, the dictionary building process to support and facilitate text analytics is an artifact that needs to be built and evaluated.

Tightly aligned with, and often subsumed within design science research is research on design process models. Prior research has proposed many design process models (see Alter, 2013; Cole, Purao, Rossi, & Sein, 2005; Eekels & Roozenburg, 1991; Gleasure, Feller & O'Flaherty, 2012; March & Smith, 1995; Nunamaker, Chen, & Purdin, 1991; Offermann, Levina, Schönherr, & Bub, 2009; Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007; Takeda, Veerkamp, & Yoshikawa, 1990; Vaishnavi & Kuechler, 2015). Despite the differences, all previously referenced design process models include two activities, design and evaluation. One widely-adopted model (Peffers et al., 2007) divides the design process into six activities: 1) problem identification and motivation; 2) define the objectives for a solution; 3) design and development; 4) demonstration; 5) evaluation; 6) communication.

Although design process models provide some general descriptions of the process of conducting design science research, they do not 'unpack' the specific steps, 'design', nor do they provide practical guidelines on how to design. Our aim in this paper is to reveal the dictionary building process and to provide researchers with practical guidelines for building a dictionary, which, obviously, cannot be fulfilled by proposing one general step, '*design a dictionary*'. Design science and design process models do bring several advantages. First, despite the lack of practical guidelines, the design process models do describe a complete high-level process for completing a design science research project which provides us a starting point for developing the dictionary building process. Second, the design process models emphasize the importance of evaluation, which is overlooked by most prior dictionary building research (exceptions being Grimmer & Stewart, 2013 and Krippendorff, 2004 who have proposed several preliminary validation criteria). In this paper, we take a step forward to uncover the '*design*' in the design process. We

ISSN: 1941-6679-On-line Copy

follow March & Smith (1995) to develop a process model for semi-automatic dictionary building with the focus on *design* and *evaluation*.

METHOD

To accomplish our goal of designing a process model for semi-automatic dictionary building. we followed the inductive consensus-building approach used by Peffers et al. (2007) in developing the *Design Science Research Process Model*. Specifically, we examined prior research where dictionaries were built to determine and infer the appropriate elements and steps required in dictionary building. We synthesize said literature to explicate an initial set of required dictionary building steps resulting in a process model that is consistent with the existing research. Thus it would serve as a commonly accepted framework for carrying out dictionary building research.

To identify the research involving dictionary building activities, we conducted several rounds of search in *Web of Science* and *Google Scholar* using keywords, such as "*dictionary building/development/developing/construction*", "*automated/ automatic content analysis*", and "*computer-assisted content analysis*". In addition, we browsed the websites of text mining software (e.g., WordStat, LIWC, etc.), with the aim of finding existing available dictionaries and then tracing back to their sources. Following these two steps resulted in 18 initial papers. To expand our sample, we adopted a snowball sampling strategy. We reviewed the introduction and literature review sections of the 18 papers, to identify any additional related papers. Then we examined the introduction and literature review parts of newly identified papers. After several iterations of the aforementioned process, our sample consisted of 82 papers. We reviewed the papers and filtered our sample to only include research on semi-automatic dictionary building. After the filtration, 54 papers were removed from the sample (16 papers for not mentioning the dictionary building process; 21 papers on manual or automatic dictionary building; 5 papers on general discussion; 12 unrelated papers). In total, our final sample includes 28 papers that contain some aspect of a dictionary building process.

Although none of the 28 papers provides a normalized comprehensive dictionary building process, they do include many descriptions of portions of their dictionary developing processes. Following the inductive consensus-building approach, where possible, we analyzed the descriptions of dictionary building processes (or lack thereof) in these papers, summarized the steps adopted (see Table 3), and subsequently derived a general dictionary building process. The resulting process is described in full in the next section.

ISSN: 1941-6679-On-line Copy

Table 3. Summary of dictionary building process

Citation	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6
Aaldering & Vliegenthart (2015)	●	◉		○	○	●
Abrahamson & Eisenman (2008)	●	●	○	●		○
Albaugh et al. (2013)	●	◉		◉	○	○
Bengston & Xu (1995)	●	●		◉		●
Boritz et al. (2013)	●	●	●	●		●
Cohen (2012)	●	◉		◉		○
Debortoli et al. (2014)	●	○	○	○	○	
de-Miguel-Molina et al. (2016)	●	○		○		
Guo et al. (2016)	●	◉	○	○	○	◉
Kirilenko et al. (2012)	●	○		○		○
König & Finke (2015)	●	◉		○		○
Laver & Garry (2000)	●	○		◉		
Lesage & Wechtler (2012)	●	●	○	○	○	
Loughran & McDonald (2011)	●	○	○	○		
Martindale (1975, 1990)	●	◉		○		◉
Matthies & Coners (2015)	●	◉	○	○		
Mergenthaler (1996, 2008)	●	◉		○	○	
Opoku et al. (2006)	●	◉		○	●	
Park et al. (2009)	●	●		○		
Pennebaker et al. (2015)	●	○	○	◉	○	●
Péladeau & Stovall (2005)	●	●	◉	●	◉	●
Rooduijn & Pauwels (2011)	●	●		◉	○	●
Smith & Chang (1996)	●	◉		○		
Strapparave & Valitutti (2004)	●			◉		
Vasalou et al. (2011); Gill et al. (2011)	●	●	◉	◉	○	●
Wade et al. (1997)	●	●		○		◉
Wilson (2006)	●	○		○	○	◉
Young & Soroka (2012)	●	◉		◉	○	●

*Note: Step 1 (Objective Clarification); Step 2 (Corpus Creation); Step 3 (Pre-processing); Step 4 (Entry Identification & Categorization); Step 5 (Extension & Simplification); Step 6 (Validation);

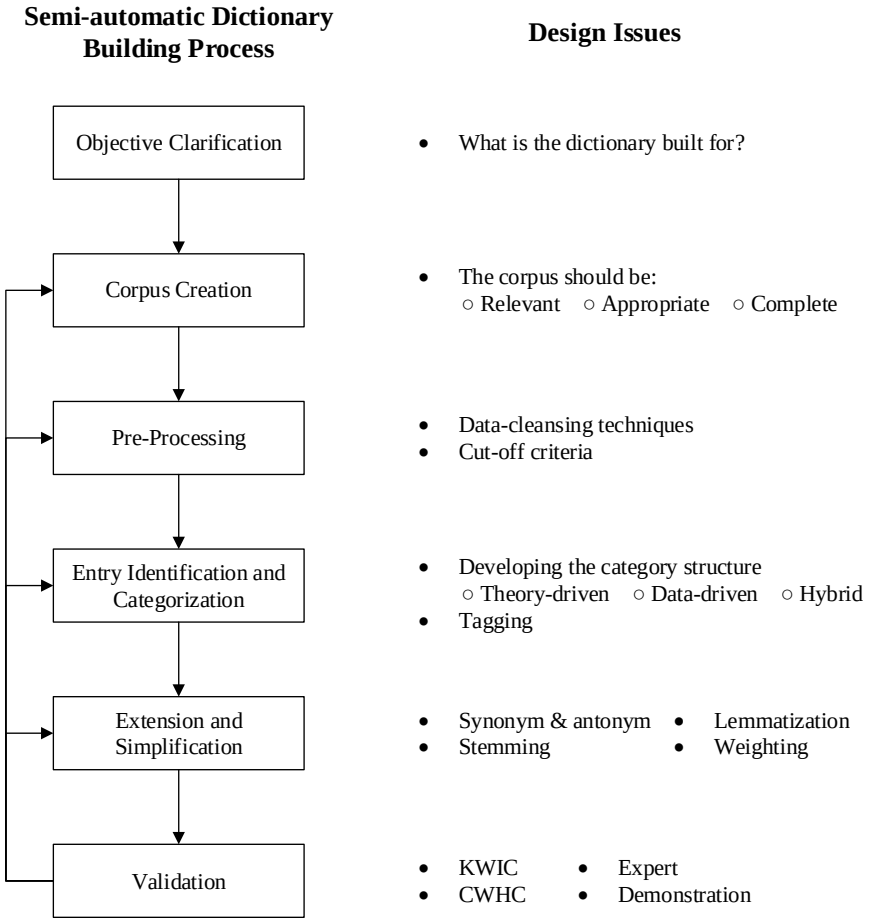
**Note: ●-sufficiently discussed; ◉-slightly discussed; ○-mentioned

ISSN: 1941-6679-On-line Copy

ARTIFACT DESCRIPTION

We name the resulting documentation the "semi-automatic dictionary building process" (S-DBP). The S-DBP includes six steps, namely, objective clarification, corpus creation, pre-processing, entry identification and categorization, extension and simplification, and validation (see Figure 1). While iteration within the steps is common, we will discuss the steps in a linear fashion.

Figure 1. The Semi-Automatic Dictionary Building Process (S-DBP)



Step 1. Objective clarification.

The dictionary building process starts with the clarification of objective. Researchers need to specify what the dictionary is being built for. For example, one can build a dictionary for theory testing, monitoring the evolution of specific topics,

or even identifying new concepts. To clarify the objectives of the dictionary, questions such as, "what is the theme of the dictionary?", "how will the dictionary be used after developed?", and "are there any appropriate and available dictionaries?" should be answered. Through answering these questions, one can confirm the necessity of building the dictionary and establish a solid basis for conducting successive dictionary building steps.

Step 2. Corpus creation.

The corpus is the set of documents from which the dictionary is developed. It usually consists of multiple documents which include rich textual contents related to the topic of the dictionary. Assembling a corpus involves selecting the right textual sources for future processing. Since the dictionary is derived from the corpus, its quality and applicable scope are directly dependent on the documents in the corpus.

Although all of the identified 28 papers provided the descriptions of their corpora (see Table 3), none of them has provided an assessment of corpus. Three features of the corpus could be considered to decide whether the corpus is "adequate". First, the corpus should be relevant. It should include the contents which are consistent with the theme of the dictionary being built. Second, the corpus should be appropriate. If the dictionary being built includes only words/phrases, the original corpus should include mainly textual contents, instead of numeric or pictorial contents. Sometimes, the dictionary needs to include more than words and phrases. For example, the LIWC 2015 can now accommodate numbers, punctuation, and even short phrases, which allows users to analyze "netspeak" language that is common in the context of online communication (e.g., Twitter and Facebook posts, text message, etc.). In the LIWC 2015, "b4" is coded as a preposition and ":)" is coded as a positive emotion word (Pennebaker et al., 2015). Third, the corpus should be complete. For example, in building a dictionary of forest values, Bengston and Xu (1995) created a corpus which includes articles by forest economists, traditional foresters, forest ecologists, landscape architects, aestheticians, environmental philosophers, environmental psychologists, Native Americans, among others. To be complete does not mean that the corpus should include every related document; instead, it means that the richness and completeness of the corpus should be adequate to support the dictionary building. The criterion of "completeness" is especially important for dictionary building where pre-specified categories are being used. If the corpus does not include entries that map to the categories, the value of the dictionary will be sub-standard.

ISSN: 1941-6679-On-line Copy

Step 3. Pre-processing.

The aim of this step is to prepare the corpus for further analysis. There are two main types of pre-processing techniques: 1) data cleansing techniques; and 2) cut-off criteria. Data cleansing techniques include: stop word removal (see Debortoli et al., 2014), unnecessary information removal (see Lesage & Wechtler, 2012; Eriksson, Jensen, Frankild, Jensen, & Brunak, 2013), reducing phrases to single words (see Gill et al., 2011; Vasalou et al., 2011), spelling correction, among others. Some of the data cleansing techniques (i.e., unnecessary information removal, spelling correction) are almost always necessary, while the others (i.e., stop word removal, reducing phrases to single words) are optional and dependent upon the goals of the research.

Researchers often determine cut-off criteria and retain/exclude entries that meet the criteria. Popular cut-off criteria include term frequency and frequency of the documents in which one entry occurs. Examples from the 28 papers include terms occurring: "more than 30 times" (Abrahamson & Eisenman, 2008), "more than 1000 times" (Guo et al., 2016), "more than 5000 times" (Boritz et al., 2013), "in less than 1% of the documents" (Lesage & Wechtler, 2012; Debortoli et al., 2014) and "in more than 5% of the documents" (Loughran & McDonald, 2011). Researchers could also set up specific cut-off criteria, such as "used by one party twice as often as by the other" (Laver & Garry, 2000) and "occur at least once in multiple corpora" (Pennebaker et al., 2015). TF*IDF is another popular cut-off criterion. TF refers to term frequency, and IDF refers to inverse document frequency. Although TF*IDF has not been used in the papers we reviewed, it is a standard way of culling words up front. The usage of this metric is based on the assumption that the more frequent a term occurs in a document, the more representative it is of the document's content yet, the more documents in which the term occurs, the less important the term is in distinguishing different documents' content from each other. So, if the purpose of the research is to distinguish between documents, as it is in classification tasks, TF*IDF is extremely important. As our review indicates, the cut-off criterion is usually an arbitrary decision made by researchers based on the scope of the corpus or a decision to follow established criteria levels from previous studies. Usually, the pre-processing is conducted with the help of text analysis or text mining software. Currently, there is much computer-aided text analysis (CATA) software can assist with the pre-processing step (for example, *WordStat* and *RapidMiner* among others). In this step, the choice of techniques is a decision that is made by researchers based on the requirement of the dictionary. Of the 28 identified papers, 11 include this step, and 17 do not.

ISSN: 1941-6679-On-line Copy

Step 4. Entry identification and categorization.

A dictionary typically includes three basic elements: the entries (words, word stems, and phrases), the categories, and the association between the entries and the categories. Categories, according to Weber (1983, p. 140) are "*a group of words [and phrases] with similar meaning and/or connotations*". In this step, there are two core activities, developing the category structure and categorizing entries. For projects that have pre-specified categories, the main activity in this step is entry categorization. For projects that do not have pre-specified categories, researchers can use several approaches (e.g., theory-driven, data-driven or hybrid) to develop the category structure. The theory-driven approach is a method where researchers develop category structures based on the related theories (see Aaldering & Vliegenthart, 2015; Abrahamson & Eisenman, 2008; Bengston & Xu, 1995; Debortoli et al., 2014; Laver & Garry, 2000; Loughran & McDonald, 2011; Opoku et al., 2006; Péladeau & Stovall, 2005; and Young & Soroka, 2012). For projects that are more exploratory in nature, category structures can be derived using a data-driven approach (see Kirilenko et al., 2012; Lesage & Wechtler, 2012). Typically, this is done with the aid of a 'topic extraction' feature within text mining software that aids in uncovering the thematic structure of the processed text. Topic extraction is usually implemented using latent semantic analysis, latent dirichlet allocation or factor analysis. The category structure could also be developed using a hybrid approach (see Boritz et al., 2013; Cohen, 2012; de-Miguel-Molina et al., 2016; Gill et al., 2011; Guo et al., 2016; Vasalou et al., 2011). In these situations, researchers usually start with the pre-specified category structures derived from theory and then modify the category structures according to the text mining results (e.g., topic extraction, etc.) during the dictionary building process. There is no superior or inferior approach, and the choice will be project dependent. For example, a theory-driven approach is more suitable for confirmatory studies (e.g., theory testing, concept identification, etc.), while the data-driven approach is more suitable for exploratory studies (e.g., theory building, concept formation, etc.).

Typically, in the semi-automatic dictionary building process, entry categorization is manually conducted by researchers, who are familiar with the theme of the dictionary, with the assistance of text analysis software. Researchers examine each entry in the list developed in step 3 and decide whether the entry should be retained and into which category the entry should be assigned. In most of the studies we reviewed, the entry identification and categorization are conducted by the single researcher. However, it can be performed by multiple researchers as well (Abrahamson & Eisenman, 2008; Cohen, 2012; Gill et al., 2011; König & Finke, 2013; Opoku et al., 2006; Pennebaker et al., 2015; Vasalou et al., 2011). In the multi-coder case, the concept of inter-coder reliability is introduced as an assessment of the word categorization (see Abrahamson & Eisenman, 2008). The

ISSN: 1941-6679-On-line Copy

result of this step is an initial dictionary which should be further modified and validated before being directly applied to analyze text documents.

Step 5. Extension and simplification.
The most common techniques are synonym and antonym extension, stemming, lemmatization and weighting. Synonym and antonym extension refers to adding synonyms (and antonyms) to the initial words in the dictionary. Sometimes, this is the major way of identifying entries (see Opoku et al., 2006). Because of the various wording preferences, different terms might be used by different authors to express the same meaning. Therefore, extending the dictionary by including synonyms (and antonyms) can, to some degree, increase the generalizability of the dictionary.

To efficiently and effectively find insights in text, dictionary entries are often reduced through stemming or lemmatization. Stemming is a more rudimentary approach where words are simply truncated. For example, the word "having" may be stemmed to "hav*". Alternatively, lemmatizing aims to retain the morphology of the word and would thus reduce "having" to "have". The choice of approach is project dependent. Stemmers are faster and simpler, but lemmatization is more accurate. In this way, the dictionary can be simplified without sacrificing accuracy and effectiveness.

Weighting means to weight terms based on their occurrence in and across documents. It is usually performed by applying the previously mentioned TF*IDF (Term Frequency-Inverse Document Frequency) weighting scheme (see Debortoli et al., 2014). Compared with synonym and antonym extension, stemming, and lemmatization, weighting is less commonly used. However, in some special cases, this technique can promote the occurrence of rare terms and discount the occurrence of more common terms (Debortoli et al., 2014; Manning, Raghavan, & Schütze, 2008). Similar to Step 4, each modification of dictionary in this step needs to be carefully examined and validated.

Step 6. Validation.
The fourth step results in an extended and simplified dictionary that should be validated before being widely applied. Of the 28 papers reviewed, 17 report some form of validation of the dictionary. As the review shows, the validation methods include key-words-in-context (KWIC) (9 papers), compare-with-human-coding (CWHC) (5 papers), expert validation (3 papers) and demonstration (2 paper). Since the same entry might have different meanings in different contexts, it is necessary to have a look at the actual usage of the entry in the corpus to determine whether the entry is the accurate indicator of the concept the researcher perceives

it to indicate. KWIC facilitates this process and is a common feature in most text mining software. In CWHC, the similarity between the automated coding results and human coding results are the primary indicator of the validity of the dictionary. The dictionary can also be validated by having a domain expert review and, if necessary, adjust the contents of the dictionary. For example, to validate the forest value dictionary, Bengston and Xu (1995) invited a landscape architect and an environmental psychologist to review the dictionary and suggest additional entries. Finally, demonstration of the use of the dictionary has been used as a method of validation. For example, Abrahamson and Eisenman (2008) applied their rational-normative dictionary to analyze the pre-designed rational and normative texts to see if the dictionary could produce results which reveal the difference between the two types of texts.

Although we illustrate the dictionary building process as a sequential step-by-step process, in reality, dictionary building is an iterative process where steps are often revisited. For example, *Validation (via KWIC or other approaches)* and *Entry Identification and Categorization* are often recurrently conducted together. If the validation indicated that the dictionary developed is not good enough, then one needs to re-think the previous steps (i.e., *Corpus creation*, *Pre-processing*, *Entry identification and categorization*, *Extension and Simplification)* to see what could be done to improve the dictionary. After being validated, the dictionary can be used to analyze the texts clarified in the first step. If one wants to use the dictionary to analyze other texts, one needs to validate the dictionary using the texts to be analyzed before actually analyzing them. Given its iterative nature, dictionary building is a time-consuming process without an objective "stopping rule" (Boritz et al., 2013). Normally, the refinement of the dictionary should be repeated until a satisfactory level of validity is achieved (Bengston & Xu, 1995). A "satisfactory level" is a rule of thumb which could be defined by researchers according to the requirements of the dictionary project. Building a comprehensive dictionary is a long-term activity which could last from months to years (Albaugh et al., 2013; Péladeau & Stovall, 2005; Pennebaker et al., 2015). However, not every dictionary is necessarily comprehensive. The scope of the dictionary is decided based on the purpose of the research. The dictionary can be used confidently as long as it is comprehensive enough to support its purpose. In next section, we will demonstrate and evaluate the S-DBP through building an environmental sustainability dictionary for the IT industry.

ISSN: 1941-6679-On-line Copy

EVALUATION

To demonstrate and evaluate the S-DBP, a 'proof of concept' is provided in this section. Proof of concept is a realization of a certain method or idea to demonstrate its feasibility, or a demonstration in principle, whose purpose is to verify that some concept or theory has the potential of being used (Gregg, Kulkarni, & Vinzé, 2001; Nunamaker et al., 1990). It has been widely used in research areas, such as engineering, business development, software development, as well as design science research (see Becker, Breuker, & Rauer, 2011; Li & Larsen, 2011; Truex, Alter, & Long, 2010). In this section, we present a 'proof of concept' for the S-DBP by following its steps to build an environmental sustainability dictionary for the IT industry. The selection of this context was shaped by our belief in the potential value of dictionary-based text analytics approach to research on environmental sustainability reporting as well as the current lack of a dictionary specialized in sustainability. We use *WordStat*, a text mining software from *Provalis Research*, to support the dictionary building process. *WordStat* has been used extensively in dictionary building related research (see Bengston & Xu, 1995; Boritz et al., 2013; de-Miguel-Molina et al., 2016; Laver & Garry, 2000; Loughran & McDonald, 2011; Opoku et al., 2006; Wilson, 2006; Young & Soroka, 2012).

Step 1: Objective clarification.

Research on environmental sustainability reporting has a long history of using a manual content analysis method based on human coding. To our knowledge, a dictionary-based text analytics approach has rarely been applied in this research area. Our aim is to build an environmental sustainability dictionary which can be used to analyze the contents of corporate sustainability reports. Since the main objective of this section is to demonstrate and evaluate the S-DBP, we limit the scope of the dictionary by focusing on IT industry and relying on data from a single year.

Step 2: Corpus creation.

Corporate sustainability reports of IT companies from the *2015 Fortune 500* were collected and used to create the corpus for our dictionary building exercise. Corporate sustainability reports include environmental sustainability contents; they are thus related. Despite the presence of some numerical data, most of the contents of corporate sustainability reports are textual data, and therefore appropriate. Corporate sustainability reports are one of the most important artifacts to communicate a company's sustainability performance to its stakeholders. Therefore, it generally includes every aspect of the company's sustainability performance and thus can be considered complete. Of the 49 IT companies included in the 2015 Fortune 500, 28 issued annual corporate sustainability reports, 10 issued

ISSN: 1941-6679-On-line Copy

online sustainability disclosures, and 11 did not disclose corporate sustainability information. To improve the corpus' relatedness, we only collect the environmental section from the CS reports and online disclosures from 2015. This resulted in 751 pages (reduced from 2,119 pages) of CS report contents and 53 pages of online disclosure contents. In total, the initial corpus consists of 38 documents (reports or online disclosures, see Appendix 1), which include 804 pages of environmental sustainability related contents.

Step 3: Pre-processing.

After importing the initial corpus into *WordStat*, we conducted two steps of pre-processing. First, two data cleansing techniques, spell check and stop word (e.g., "a", "and", "or", etc.) removal, were used. Although corporate sustainability reports and online disclosures are official publications and typically do not include spelling mistakes, it is still necessary to conduct a spell check before further analysis because the format of the textual data might change while importing the data into the text analytics software. For example, the original phrase, "*environmental sustainability*", might become "*environnmentalsustainability*" after being imported. Since these format changes influence the frequency analysis later, it is necessary to deal with them before conducting next step. The spelling check can be conducted with the help of built-in functions of *WordStat*. *WordStat* also has a built-in stop word dictionary which can be refined by researchers according to the research objective. Enabling the stop word removal function will automatically exclude the stop words from the subsequent text analysis. We used the default stopwords dictionary because it does not include sustainability-related words, thus, will not impact the text analysis later. Second, the cut-off criteria were applied. After data cleansing, the corpus contained 9,832 unique words (246,870 words before deduplication). We considered both words and phrases to be potential entries in our dictionary because, compared to single words, phrases are more context-resistant. After applying the cut-off criterion of "occurring in no less than 10 (around 25% of) documents", 1,337 words were retained. After applying the cut-off criterion of "occurring in no less than 10 documents with max words of 3", 157 phrases were obtained from the corpus.

Step 4: Entry identification & categorization.

We follow a theory-driven method to develop the category structure. Specifically, we adapted the environmental sustainability categories of the GRI G4 reporting framework to support the entry categorization. This approach is consistent with many studies on corporate sustainability reporting (see Bonilla-Priego, Font, & del Rosario Pacheco-Olivares, 2014; Delai & Takahashi, 2013; de Grosbois, 2015; Gill, Dickinson, & Scharl, 2008). The GRI G4 environmental sustainability framework divides corporate environmental sustainability into twelve related categories. We

ISSN: 1941-6679-On-line Copy

removed the *Products & Services*, *Transport*, and *Overall* categories from our dictionary structure because they partially overlapped with the eight other sustainability-categories (i.e., *Materials*, *Energy*, *Emissions*, *Water*, *Biodiversity*, *Effluents & Waste*, *Compliance*, *Environmental Grievance Mechanisms*). For example, the GRI asks corporations to report '*Products and Services*' from the perspectives of materials, energy, emissions, etc. Thus, the reported contents for the category, *Products and Services*, often co-exist in other sustainability categories. A similar situation can be found for the removed *Transport* and *Overall* categories. This can cause problems in any analysis that is done. For example, if we categorize the word, "energy", into the *Energy* category, then the software would automatically count the "energy" occurring in the section of *Products and Services*, and in this way, the analysis result of *Products and Services* would be invalid. However, the problem of overlapping is not unsolvable. To analyze the sustainability contents of the *Products and Services*, one could use two dictionaries (one for the *Products and Services* and one for *Energy*, *Emissions*, and so on) and examine the co-occurrence of the words in the two dictionaries. We also removed the *Supplier Environmental Assessment* from our categories because, 1) from the data perspective, it also partially overlaps with the eight sustainability-focused categories, and 2) from the theory perspective, its main focus is on the approach of supplier management, and not on the sustainability performance of supplier. Of the eight remaining categories, we extended the scope of *Compliance* from non-compliance behavior to both mandatory compliance (e.g., compliance with environmental laws and regulations) and voluntary compliance (e.g., voluntarily pursuit of environmental certifications) behaviors.

The first author then manually reviewed the words and phrases retained after step 3, aiming to identify environmental sustainability-related entries and categorize them into the eight categories identified above. To properly assess the retained words, one needs to be aware of acronyms, word co-occurrence, context, and word forms. For example, "led" could mean "LED lighting", but it is also the past participle of "lead". Combinations of a specific word with other words can introduce different meanings. For example, "efficiency" by itself appears to be a sustainability-related word. However, in CSR it typically is paired with other words such as "energy efficiency" and "water efficiency". The meaning of words are often contextualized. For example, "scope", at first glance, is not related to any of the eight categories. However, in the context of sustainability reporting, it is a specific word that being used in the section of *Emission* as "scope 1/2/3 emission". Finally, different forms of the same word may have different meanings. For example, in the sustainability context, "cells" is always used as "fuel cells" or "solar cells", and thus would be placed into the *Energy* category, while "cell" is always used as "cell phone" and is not a sustainability-related word.

Each environmental sustainability entry was identified, assessed and categorized based on its examination using the keywords in context (KWIC) approach. This initial attempt resulted in a dictionary containing 165 entries. Since only two entries were identified and categorized into the category, *Environmental Grievance Mechanisms (EGM)*, we combined it with *Compliance*.

Step 5: Extension & simplification.

For the words in the initial dictionary, we examined their synonyms and antonyms, which also occur in the documents, to see whether they should be included in the dictionary. Similar to the initial coding, this step was also guided by the category schema and with the help of KWIC. One thing to notice is that the cut-off criteria are not applied to the synonyms. This step generated 33 new words. We did not conduct stemming or lemmatization because sometimes different tenses of a word will have different meanings. Finally, since this was the first step to build an environmental sustainability dictionary, we did not weight the entries.

Step 6: Validation.

We conducted four rounds of validation of the dictionary. KWIC method was used in the first round, where we designed a task of re-coding the previously identified entries into the dictionary categories. A trained doctoral student (coder 1, who is familiar with corporate sustainability topics and concepts) and the second author (coder 2) conducted this task. The coders were instructed to categorize the identified entries resulting from steps 4 & 5 into the seven environmental sustainability categories. They were provided an introduction to the GRI G4 environmental sustainability framework as well as written document explaining each category. Coders used the KWIC function of *WordStat* in performing the assigned task. Both coders were unaware of the original categorization of the entries. The inter-reliability is shown in Table 4 below.

ISSN: 1941-6679-On-line Copy

Table 4. Inter-coder reliability of the entry categorization

No.	Category	Number of Entries	Reliability*	
			Coder 1	Coder 2
1	MATERIALS	34	0.79	0.79
2	ENERGY	63	0.92	0.97
3	WATER	6	1.00	1.00
4	BIODIVERSITY	5	0.80	0.80
5	EMISSIONS	16	1.00	1.00
6	EFFLUENTS & WASTE	38	0.97	0.79
7	COMPLIANCE & EGM	36	0.89	0.81
	All Entries	**198**	**0.91**	**0.87**

*Scale of the inter-coder reliability: *0.21-0.40 (Fair); 0.41-0.60 (Moderate); 0.61-0.80 (Substantial); 0.81-1.00 (Almost Perfect)* (Cohen, 1960; Landis & Koch, 1977).

As shown in Table 4, the overall inter-rater reliability is almost perfect (i.e., 0.91 for coder 1 and 0.87 for coder 2). In the second round of validation, an expert on corporate sustainability (the fourth author) re-examined every entry coded differently from coder 1 or coder 2 with the assistance of KWIC and discussed the entry context with the two coders. The dictionary was refined based on the discussion. The final dictionary included 192 words and phrases, a portion of which are shown in Table 5.

Table 5. Dictionary of environmental sustainability for IT industry (sample)

No.	Category	Entries
1	MATERIALS	chemicals, conflict free sourcing, Congo, DRC, hazardous materials, hazardous substances, material, materials, mineral, minerals, paper, plastic, plastics, sourcing, substance, substances, tantalum, tin, tungsten
2	ENERGY	battery, cells, clean energy, cooling, electricity, energy, energy consumption, energy efficiency, farm, fuels, gasoline, grid, heating, HVAC, kilowatt, kilowatts, KW, KWH, LED lighting, lighting, solar
3	WATER	Irrigation, water, water consumption, water conservation, water usage, water management
4	BIODIVERSITY	Forest, forests, trees, wildlife

5	EMISSIONS	Air emissions, carbon dioxide, carbon emissions, dioxide, emission, emissions, GHG emissions, greenhouse, GHG, scope, greenhouse gas
6	EFFLUENTS & WASTE	Batteries, composting, discharge, discharged, effluent, effluents, electronic waste, end of life, hazardous waste, landfill, recyclability, recyclable, recycled, recycling programs, reuse, reusing, scrap, solid waste, waste, wastewater, waste management, waste reduction
7	COMPLIANCE & EGM	Agencies, compliance, certification, complying, EICC, environmental laws, energy star, greenhouse gas protocol, ISO, laws, laws and regulations, LEED, legal, legislation, OECD, regulations, restriction of hazardous, ROHS, violations, violation

For the third round of validation, we used the CWHC method to assess the performance of the developed dictionary. From our initial sample, we selected the organizations that had issued sustainability reports across multiple years. This filtering resulted in 22 companies being selected. Considering that our corpus includes mainly the sustainability reports issued after 2009 and the potential evolution of sustainability-terminology, we adopted a cut-off criteria of "after 2009" here to ensure the validity the dictionary. For each organization, we randomly chose a year after 2009 and collected the associated sustainability report. We purposefully avoided using any sustainability reports inform our dictionary building task in this validation stage. Ultimately, 22 reports were collected (see Appendix 1). We randomly selected 15 paragraphs from the environmental section of each report. In total, we collected 330 paragraphs. Then, using the dictionary and associated categories, we determined the major topic of each paragraph based on the highest frequency count of dictionary words. For example, if a paragraph had 5 occurrences of 'energy' words/phrases and 3 occurrences of 'emissions' words/phrases, the paragraph would get coded as 'energy'. Two independently trained coders (the doctoral student in KWIC and the expert in our second round of validation) then manually coded each of the 330 paragraphs into one of the dictionary categories. We added two extra categories, *Multiple Topics* and *No Specific Topic*, to represent paragraphs that the software could not determine a major topic (e.g., a paragraph with 5 occurrences of 'energy' words/phrases and 5 occurrences of 'emissions' words/phrases) and paragraphs that do not include any entries that exist in the dictionary. We consider *multiple topic* paragraphs to match if the topic identified by a coder is the same as one of the multiple topics decided by software. *No Specific Topic* paragraphs are counted as a match if the coder also could not identify a topic based on the provided categories. Results are presented in Table 6.

Table 6. Reliability of CWHC

No.	Category	Automated Coding	Reliability*	
			Coder 1	Coder 2
1	MATERIALS	27	0.70	0.74
2	ENERGY	105	0.94	0.86
3	WATER	23	0.87	0.91
4	BIODIVERSITY	3	1.00	1.00
5	EMISSIONS	38	0.97	0.82
6	EFFLUENTS & WASTE	52	0.96	0.81
7	COMPLIANCE & EGM	30	0.93	0.93
8	MULTIPLE TOPICS	31	1.00	0.90
9	NO SPECIFIC TOPIC	21	0.57	0.57
	All Paragraphs	**330**	**0.91**	**0.83**

*Scale of the inter-coder reliability: *0.21-0.40 (Fair); 0.41-0.60 (Moderate); 0.61-0.80 (Substantial); 0.81-1.00 (Almost Perfect)* (Cohen, 1960; Landis & Koch, 1977).

The average reliability between automated coding and human coding is 0.87 (specifically, coder 1 is 0.91 with the automated approach and coder 2 is 0.83 with the automated approach). Overall, this falls in the 'almost perfect' reliability category according to Cohen (1960) and Landis & Koch (1977). The reliability for *No Specific Topic* paragraphs is only 0.57, which is at a moderate level. This means that, of the 21 paragraphs coded by software as *No Specific Topic*, 9 were coded as a sustainability topic by the human coders. This is a possible indication that more entries need to be added to the dictionary to identify the sustainability topics. The cut-off criteria we adopted in the pre-processing step of dictionary building might be responsible for this result. Overall, based on the multi-stage validation process, we believe that following the S-DBP has resulted in a dictionary that is valid.

The fourth type of validation is a demonstration. The purpose of the demonstration is to show how the resulting dictionary can be used in an analysis of environmental sustainability for technology companies. Because of the nascent stages of dictionary development, we are cautious about drawing any decisive conclusions from the results reported below.

For the demonstration, we collected 39 corporate sustainability reports of 13 Fortune 500 IT companies for the years 2009, 2012 and 2015 (see Appendix 1). We sampled over three years to determine if the contents of the environmental sustainability sections of the reports changed over time (based on the categories of the dictionary). Using *WordStat,* we detected all the words/phrases from the

dictionary in the environmental sustainability sections of the reports and generated a contingency table showing the number of words in each of the dictionary categories across the year of publication (see Table 7 below). This data can then form the basis of analysis that adds insight into how the different topics (represented by categories) of environmental sustainability ebb and flow across time as reported in their formal reports. From the table, it is clear that more environmental sustainability words are being detected in the 2012 and 2015 reports than in the 2009 reports and that the most common category across years is *Energy* followed by *Effluents & Waste, Emissions* and *Materials. Biodiversity* has the least amount of words being detected. While these are definitive statements, they need to be considered knowing that there is not an even distribution of dictionary words across environmental sustainability categories.

Table 7: Environmental sustainability words in corporate sustainability reports

Category	**Year**			**Total**
	2009	**2012**	**2015**	
MATERIALS	667	874	909	**2450**
ENERGY	1588	2455	2330	**6373**
WATER	350	612	540	**1502**
BIODIVERSITY	56	40	79	**175**
EMISSIONS	1027	1323	1279	**3629**
EFFLUENTS & WASTE	1093	1457	1228	**3778**
COMPLIANCE & EGM	401	686	672	**1759**
Total	**5182**	**7447**	**7037**	

Because the outcome of the application of text mining is often a contingency table, it is typical to report results using correspondence analysis (CA). CA is a method that allows the graphical representation of contingency table data in low-dimensional space (Greenacre, 2007). CA has been successfully used in a variety of domains including marketing (Inman, Shankar, & Ferraro, 2004), tourism management (Opoku, 2009; Pitt, Opoku, Hultman, Abratt, & Spyropoulou, 2007; Rojas-Mendez & Hine, 2016), teaching and learning (Askell-Williams & Lawson, 2004) among others.

The first step in CA is to test the "homogeneity assumption" (Greenacre, 2007) about whether significant differences exist between the different years' corporate sustainability reports in terms of the amount of environmental sustainability words

ISSN: 1941-6679-On-line Copy

and phrases in the dictionary categories. This assumption is tested using the chi-square statistic and is reported in Table 8. Given the chi-square value of 77.914, we can reject the hypothesis and conclude that real differences exist between the different years' report contents with regards to the seven different sustainability categories. Stated another way we can say that that there is a statistical dependence between the rows and columns of the contingency table shown in table 8.

Table 8. Summary statistics

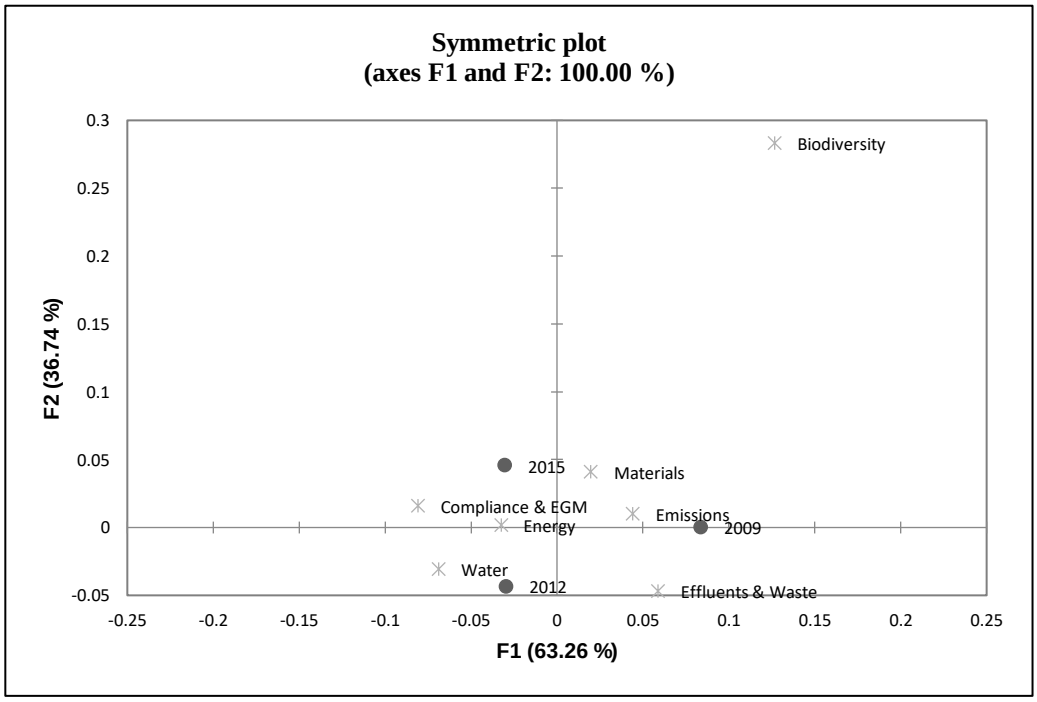
Dimensional Representation	**Eigenvalues/ Inertia**	**Chi Square**	**Percentage of Inertia**	**Cumulative Percentage**
1	.003		.633	.633
2	.001		.367	1.00
Total	**.004**	**77.914[a]**	**1.000**	**1.000**

[a] p<.0001; df 12

Note there are four- dimensions listed in the summary table. The number of dimensions in CA will be (y-1) where y is the minimum number of columns or rows in the contingency table. In our model, the first dimension explains 63% of the total inertia in the model and the second dimension explains 37%. While there are several types of CA maps available, Greenacre states that "the symmetric map is the best default map to use" (Greenacre, 2007, p. 267). The symmetric map typically provides a 'nicer-looking' representation than the asymmetric approach which often compresses the primary coordinates of the row profiles towards the center of the map to allow the display of the extreme vertices of the column profiles (essentially creating a map that is more difficult to visualize than a symmetric map). The CA map of the contents of the years' reports as detected by the sustainability dictionary is shown in Figure 2 below.

ISSN: 1941-6679-On-line Copy

Figure 2. CA map of yearly corporate sustainability reports to sustainability categories



The point at which the axes cross represents the average yearly profile of environmental sustainability topics. If we look primarily at the horizontal axis, which in CA explains more of the variance than the vertical axis, we see that the yearly profiles are the most different between {2012; 2015} and 2009 as the horizontal distance between these years is the greatest. By envisioning a line emanating from the average profile location through a category data point and then assessing the distance from the resulting line to a year profile point, we can estimate the relative proportion of said category to the yearly profiles. So, for example, the 2012 reports have proportionally more entries in *Water* than in the other two report years. Similarly, the 2012 and 2015 reports have proportionally more entries in *Energy* and *Compliance & EGM* than the 2009 reports. There are more proportional entries in *Materials* and *Biodiversity* in 2015 and 2009 than there are in 2012. *Emission* entries are proportionally higher in 2009 than in 2012 and 2015. Finally, there are proportionally more entries in *Effluents & Waste* in 2009 and 2012 reports than there are in 2015 reports.

ISSN: 1941-6679-On-line Copy

DISCUSSION

In this paper, we developed a semi-automatic process model for dictionary building. The development of the process model is well-grounded in existing literature and can be used by further research on designing, developing and applying dictionaries in text analytics projects. While this paper represents a unique effort to formally define a dictionary building process model, three cautionary points should be considered. First, researchers should be aware that the S-DBP is *not the only* appropriate methodology for developing a dictionary. As discussed in section 2.2, there are several other approaches (i.e., manual and automatic) to develop a dictionary. Second, there is no need to adopt the S-DBP as a rigid orthodoxy. The S-DBP aims to provide prescriptive guidelines, rather than impose requirements. The S-DBP can be adapted and customized for individual research projects. Finally, as stated earlier, "*computer-based investigation is no better than the dictionaries it employs. If the dictionaries are silly, the study itself will be foolish*" (Hart, 1984, p.15).

Properly Positioning the Value of the S-DBP

The importance of a normalized dictionary building process is emphasized in this paper. However, in the academic community, the value of dictionary-based text analytics is not without controversy. Some criticize building a new dictionary for its high cost, low efficiency, low generalizability and high uncertainty and propose non-dictionary-based automated text analysis (or text mining) as an alternative (see Landmann & Zuell, 2008; Wiedemann, 2013). Others recognize that once a dictionary has been built, it offers low marginal cost, high capability, prevision and high consistency (see Boritz et al., 2013; Cohen, 2012). Here, following Grimmer and Stewart (2013), we believe that there is no globally best method for automated text analysis. Different data sets and different research questions necessitate different analysis methods. While use cases of the dictionary-based method are abundant (as detailed in Guo et al., 2016) researchers need to carefully consider effective ways to apply the method.

Decision-making within the S-DBP

Semi-automatic dictionary building is an iterative process which involves both computer computations and human interventions. During the process, researchers need to make many decisions (e.g., which documents should be included in the corpus, should stemming be used, which cut-off criteria should be applied, etc.) based on their own expertise. Reviewing the current literature suggests that often

ISSN: 1941-6679-On-line Copy

these decisions are either arbitrary or at a minimum, not sufficiently justified. Our review found that most of the prior studies did not disclose the dictionary building processes adequately. The S-DBP partly addresses this problem by providing some general guidelines on how to make decisions during the dictionary building process. The result of each decision could impact the validity of the dictionary. For example, by applying a cut-off criterion, one risks losing some potentially important dictionary entries. To date, no research has examined the impacts of these decisions on the validity of the dictionary, nor the possible avenues to neutralize the impacts. We encourage researchers to disclose, or better justify, all the decisions they make and the underlying rationale to improve the transparency of the processes they adopted to build their dictionaries.

Applying the Concept of 'Confidence Level' to Dictionary Building

Given the complexity and variability of word meanings, no matter how careful one is in the selection of words and phrases to measure a specific dimension, it is likely that the inclusion of some entries will result in categorization errors or false positives (Péladeau & Stovall, 2005). Dictionary builders sometimes find themselves in a dilemma, where they have to balance the generalizability against the validity of the dictionary. For example, consider adding the word *power* into the *Energy* category in the sustainability dictionary. The word *power* occurs 100 times in the corpus. The KWIC examination indicates that, of the 100 occurrences of *power*, it is used to indicate *electricity* 95 times and *political strength* 5 times. We know that in the context of environmental sustainability, *power* is widely used as an indicator of the concept of energy, and including *power* could improve the generalizability of the dictionary. However, we also notice the loss of validity of the dictionary. In this situation, should one include the word *power* in the category *Energy* of the dictionary? What if the *power* is used to indicate *electricity* 80 times and *political strength* 20 times? To address this issue, we propose using a '*confidence level (CL)*' which can be calculated as follows (for word *x*):

$$CL_x = \frac{Times\ of\ true\ positives}{Times\ of\ occurrence}$$

True refers to the concept-congruent usage of the words. In the first example, the confidence level of the word, *power*, is 95% (or 0.95). The general confidence level of one dictionary can be the average of the confidence levels of entries in the dictionary. The concept of confidence level has potential to neutralize the controversy between proponents and critics of the dictionary-based method. Instead of criticizing or justifying the method, it provides another mechanism to assess the validity of a dictionary. Researchers need to apply a CL that they are satisfied with

ISSN: 1941-6679-On-line Copy

for their research project and domain of study. Note that appropriate CLs could vary across different domains. Future research could examine the impacts of different CL requirements on the effectiveness of a dictionary and determine a commonly accepted domain-dependent threshold value. We believe that CL could play an important role in future research on normalizing the dictionary building process.

CONCLUSION

In this paper, we developed a normalized process model for semi-automatic dictionary building. Positioning this paper in the *Design Science Research Knowledge Contribution Framework* proposed by Gregor and Hevner (2013), we believe that this paper has presented an *improvement-type* contribution (i.e., develop new solutions for known problems) as it explores *how to design* in the context of dictionary building. However, the inadequacy of extant design processes revealed in this paper still raise the requests for design science researchers to pay attention to this problem. Future design science research could develop process models or guidelines for each step defined by the extant design process.

This paper has many contributions. First, although research on dictionary building already exists, none of them has proposed a normalized dictionary building process. The S-DBP presented in this paper addresses this current research gap. Second, to demonstrate and evaluate the S-DBP, we built an initial environmental sustainability dictionary for the IT industry. To our knowledge, it is the first dictionary developed for the environmental sustainability of IT companies. Although this dictionary is only an initial version and still need further modifications, we do believe that the development of such dictionary will promote the adoption of an automated text analysis method in corporate sustainability area and. Third, we extend the application of design science into the text analytics domain. As far as we know, this is the first paper which addresses the problem in dictionary building process using design science research method.

This paper is not without limitations. Due to the limitation of scope, we cannot provide detailed discussions for every possible decision researcher may confront in the dictionary building process. Moreover, the development of the environmental sustainability dictionary is more a demonstration than an evaluation of the S-DBP. However, we do believe that the S-DBP could provide a nominal process for conducting dictionary building research, as well as offer a mental model for the presentation of research outcomes. Since we adopted a consensus-building method to design the S-DBP, it is inherently consistent with the prior studies on which it is based.

ISSN: 1941-6679-On-line Copy

REFERENCES

Aaldering, L., & Vliegenthart, R. (2016). Political leaders and the media. Can we measure political leadership images in newspapers using computer-assisted content analysis?. *Quality & Quantity, 50*(5), 1871-1905.

Abrahamson, E., & Eisenman, M. (2008). Employee-management techniques: Transient fads or trending fashions?. *Administrative Science Quarterly, 53*(4), 719-744.

Albaugh, Q., Sevenans, J., Soroka, S., & Loewen, P. J. (2013). *The Automated Coding of Policy Agendas: A Dictionary-Based Approach*. Paper presented at the 6th Annual Comparative Agendas Conference, Atnwerp, Beligum.

Alter, S. (2013). Work system theory: Overview of core concepts, extensions, and challenges for the future. *Journal of the Association for Information Systems, 14*(2), 72-121.

Askell-Williams, H., & Lawson, M. J. (2004). A correspondence analysis of child-care students' and medical students' knowledge about teaching and learning. *International Education Journal, 5*(2), 176-204.

Becker, J., Breuker, D., & Rauer, H. P. (2011). On guidelines for representing business models – A design science approach. In *Proceedings of the 17th Americas Conference on Information Systems* (Paper 96). Detroit, MI, USA.

Bengston, D. N., & Xu, Z. (1995). *Changing National Forest Values: a content analysis* (Research Paper NC-323). Retrieved from United States Department of Agriculture, Forest Service, Northern Research Station Website: http://www.nrs.fs.fed.us/pubs/rp/rp_nc323.pdf

Bonilla-Priego, M. J., Font, X., & del Rosario Pacheco-Olivares, M. (2014). Corporate sustainability reporting index and baseline data for the cruise industry. *Tourism Management, 44*, 149-160.

Boritz, J. E., Hayes, L., & Lim, J. H. (2013). A content analysis of auditors' reports on IT internal control weaknesses: The comparative advantages of an

ISSN: 1941-6679-On-line Copy

automated approach to control weakness identification. *International Journal of Accounting Information Systems, 14*(2), 138-163.

Brier, A., & Hopp, B. (2011). Computer assisted text analysis in the social sciences. *Quality & Quantity, 45*(1), 103-128.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Cohen, S. J. (2012). Construction and preliminary validation of a dictionary for cognitive rigidity: Linguistic markers of overconfidence and overgeneralization and their concomitant psychological distress. *Journal of Psycholinguistic Research, 41*(5), 347-370.

Cole, R., Purao, S., Rossi, M., & Sein, M. (2005). Being proactive: Where action research meets design research. In *Proceedings of the 26th International Conference on Information Systems* (Paper 27). Las Vegas, NV, USA.

de Grosbois, D. (2015). Corporate social responsibility reporting in the cruise tourism industry: A performance evaluation using a new institutional theory based model. *Journal of Sustainable Tourism, 24*(2), 245-269.

Debortoli, S., Müller, O., & vom Brocke, J. (2014). Comparing business intelligence and big data skills. *Business & Information Systems Engineering, 6*(5), 289-300.

Delai, I., & Takahashi, S. (2013). Corporate sustainability in emerging markets: Insights from the practices reported by the Brazilian retailers. *Journal of Cleaner Production, 47*, 211-221.

de-Miguel-Molina, B., Chirivella-González, V., & García-Ortega, B. (2016). Corporate philanthropy and community involvement. Analysing companies from France, Germany, the Netherlands and Spain. *Quality & Quantity, 50*(6), 2741-2766.

Eekels, J., & Roozenburg, N. F. (1991). A methodological comparison of the structures of scientific research and engineering design: their similarities and differences. *Design Studies, 12*(4), 197-203.

Eriksson, R., Jensen, P. B., Frankild, S., Jensen, L. J., & Brunak, S. (2013). Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text. *Journal of the American Medical Informatics Association, 20*(5), 947-953.

Gill, A. J., Vasalou, A., Papoutsi, C., & Joinson, A. N. (2011). Privacy dictionary: A linguistic taxonomy of privacy for content analysis. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 3227-3236). Vancouver, BC, Canada.

Gill, D. L., Dickinson, S. J., & Scharl, A. (2008). Communicating sustainability: A web content analysis of North American, Asian and European firms. *Journal of Communication Management, 12*(3), 243-262.

Gleasure, R., Feller, J., & O'Flaherty, B. F. (2012). Procedurally transparent design science research: A design process model. In *Proceedings of the 33rd International Conference on Information Systems* (Track 19, Paper 10). Orlando. FL, USA.

Gregg, D. G., Kulkarni, U. R., & Vinzé, A. S. (2001). Understanding the philosophical underpinnings of software engineering research in information systems. *Information Systems Frontiers, 3*(2), 169-183.

Greenacre, M. (2007). *Correspondence Analysis in Practice (Second Edition)*. Boca Raton, FL: Chapman & Hall/CRC.

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly, 37*(2), 337-355.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267-297.

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big Social Data analytics in journalism and mass communication comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism & Mass Communication Quarterly, 93*(2), 332-359.

Hart, R. P. (1984). *Verbal style and the presidency: A computer-based analysis*. New York, NY: Academic Press.

  ISSN: 1941-6679-On-line Copy

Hart, R. P. (2000). *DICTION 5.0: The text analysis program*. Thousand Oaks, CA: Sage-Scolari.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly, 28*(1), 75-105.

Hiller, J. H., Marcotte, D. R., & Martin, T. (1969). Opinionation, vagueness, and specificity-distinctions: Essay traits measured by computer. *American Educational Research Journal, 6*(2), 271-286.

Inman, J. J., Shankar, V., & Ferraro, R. (2004). The roles of channel-category associations and geodemographics in channel patronage. *Journal of Marketing, 68*(2), 51-71.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing an introduction to natural language processing, computational linguistics, and speech*. Upper Saddle River, NJ: Prentice Hall.

Kirilenko, A., Stepchenkova, S., Romsdahl, R., & Mattis, K. (2012). Computer-assisted analysis of public discourse: A case study of the precautionary principle in the US and UK press. *Quality & Quantity, 46*(2), 501-522.

König, T., & Finke, D. (2015). Legislative governance in times of international terrorism. *Journal of Conflict Resolution, 59*(2), 262-282.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology (Second Edition)*. Thousand Oaks: Sage.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Landmann, J., & Zuell, C. (2008). Identifying events using computer-assisted text analysis. *Social Science Computer Review, 26*(4), 483-497.

Laver, M., & Garry, J. (2000). Estimating policy positions from political texts. *American Journal of Political Science, 44*(3), 619-634.

Lesage, C., & Wechtler, H. (2012). An inductive typology of auditing research. *Contemporary Accounting Research*, *29*(2), 487-504.

  ISSN: 1941-6679-On-line Copy

Li, J., & Larsen, K. (2011). Establishing nomological networks for behavioral science: A natural language processing based approach. In *Proceedings of the 32nd International Conference on Information Systems* (Track 16, Paper 24). Shanghai, China.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35-65.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems, 15*(4), 251-266.

March, S. T., & Storey, V. C. (2008). Design science in the information systems discipline: An introduction to the special issue on design science research. *MIS Quarterly*, *32*(4), 725-730.

Martindale, C. (1975). *Romantic progression: The psychology of literary history*. Washington, DC: Hemisphere.

Martindale, C. (1990). *The clockwork muse: The predictability of artistic change*. New York, NY: Basic Books.

Matthies, B., & Coners, A. (2015). Computer-aided text analysis of corporate disclosures-demonstration and evaluation of two approaches. *The International Journal of Digital Accounting Research, 15*, 69-98.

Mergenthaler, E. (1996). Emotion-abstraction patterns in verbatim protocols: A new way of describing psychotherapeutic processes. *Journal of Consulting and Clinical Psychology*, *64*(6), 1306-1315.

Mergenthaler, E. (2008). Resonating minds: A school-independent theoretical conception and its empirical application to psychotherapeutic processes. *Psychotherapy Research*, *18*(2), 109-126.

Morris, R. (1994). Computerized content analysis in management research: A demonstration of advantages & limitations. *Journal of Management, 20*(4), 903-931.

  ISSN: 1941-6679-On-line Copy

Nunamaker Jr., J. F., Chen, M., & Purdin, T. D. M. (1991). Systems development in information systems research. *Journal of Management Information Systems, 7*(3), 89-106.

Offermann, P., Levina, O., Schönherr, M., & Bub, U. (2009). Outline of a design science research process. In *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology* (Article 7). Philadelphia, PA, USA.

Opoku, R., Abratt, R., & Pitt, L. (2006). Communicating brand personality: Are the websites doing the talking for the top South African Business Schools?. *Journal of Brand Management, 14*(1), 20-39.

Opoku, R. A. (2009). Mapping destination personality in cyberspace: An evaluation of country web sites using correspondence analysis. *Journal of Internet Commerce, 8*(1-2), 70-87.

Park, J. R., Lu, C., & Marion, L. (2009). Cataloging professionals in the digital environment: A content analysis of job descriptions. *Journal of the American Society for Information Science and Technology*, *60*(4), 844-857.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems, 24*(3), 45-77.

Péladeau, N., & Stovall, C. (2005). *Application of Provalis Research Corp.'s statistical content analysis text mining to airline safety reports*. Retrieved from Flight Safety Foundation Website: http://flightsafety.org/files/Provalis text mining report.pdf

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin.

Pitt, L. F., Opoku, R., Hultman, M., Abratt, R., & Spyropoulou, S. (2007). What I say about myself: Communication of brand personality by African countries. *Tourism Management, 28*(3), 835-844.

Rojas-Mendez J., & Hine M. J. (2016). South American countries' positioning on personality traits: Analysis of 10 national tourism websites. *Journal of*

Vacation Marketing. Advance online publication. Retrieved from: http://journals.sagepub.com/doi/full/10.1177/1356766716649227

Rooduijn, M., & Pauwels, T. (2011). Measuring populism: Comparing two methods of content analysis. *West European Politics*, *34*(6), 1272-1283.

Schrodt, P. A., & Gerner, D. J. (2012). *Analyzing International event data: a handbook of computer-based techniques*. Retrieved from: http://parusanalytics.com/eventdata/papers.dir/AIED.Preface.pdf

Schwartz, H. A., & Ungar, L. H. (2015). Data-driven content analysis of social media: A systematic overview of automated methods. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 78-94.

Simon, H. A. (1996). *The Sciences of the Artificial (3rd Ed.)*. Cambridge, Massachusetts; London, England: The MIT Press.

Smith, J. R., & Chang, S. F. (1996). *Searching for images and videos on the world-wide web* (Technical Report No. #459-96-25). Retrieved from: https://pdfs.semanticscholar.org/5117/e9f03c03659404ce5dce1d49632b680efad8.pdf

Stone, P. J., Dumphy, D. C., & Ogilvie, D. M. (1966). *The General Inquirer: A computer approach to content analysis*. Cambridge, MA: MIT Press.

Strapparava, C., & Valitutti, A. (2004). *WordNet-Affect: An affective extension of WordNet*. Paper presented at the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal.

Takeda, H., Veerkamp, P., & Yoshikawa, H. (1990). Modeling design processes. *AI Magazine, 11*(4), 37-48.

Truex, D., Alter, S., & Long, C. (2010). Systems analysis for everyone else: Empowering business professionals through a systems analysis method that fits their needs. In *Proceedings of the 18th European Conference on Information Systems* (Paper 4). Pretoria, South Africa.

Vaishnavi, V. K., & Kuechler, W. (2015). *Design Science Research Methods and Patterns: Innovating Information and Communication Technology (Second Edition)*. Boca Raton, Florida: CRC Press.

  ISSN: 1941-6679-On-line Copy

Vasalou, A., Gill, A. J., Mazanderani, F., Papoutsi, C., & Joinson, A. (2011). Privacy dictionary: A new resource for the automated content analysis of privacy. *Journal of the American Society for Information Science and Technology, 62*(11), 2095-2105.

Wade, J. B., Porac, J. F., & Pollock, T. G. (1997). Worth, words, and the justification of executive pay. *Journal of Organizational Behavior, 18*(1), 641-664.

Weber, R. P. (1983). Measurement models for content analysis. *Quality and Quantity, 17*(2), 127-149.

Whissell, C. (1989). The dictionary of affect in language. In R. Plutchnik & H. Kellerman (Eds.), *Emotion: Theory and research* (pp. 113-131). New York, NY: Harcourt Brace.

Wiedemann, G. (2013). Opening up to big data: Computer-assisted analysis of textual data in social sciences. *Historical Social Research / Historische Sozialforschung, 38*(4), 332-357.

Wilson, A. (2006). Development and application of a content analysis dictionary for body boundary research. *Literary and Linguistic Computing, 21*(1), 105-110.

Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication, 29*(2), 205-231.

  ISSN: 1941-6679-On-line Copy

APPENDIX.

Corporate Sustainability Reports Used in the Tasks of this Research

Company	Sector *	Type **	Issued Reports	DB***	CWHC	Demonstration
Microsoft	S1	A	2003-2015	2014	2010	2009, 2012, & 2015
Oracle Corporation	S1	A	2006-2014	2014	2010	N/A
Symantec Corporation	S1	A	2008-2015	2014	2012	2009, 2012, & 2015
Salesforce.com	S1	A	2012-2014	2013 & 2014	2012	N/A
Apple	S2	A	2008-2016	2016	2014	2009, 2012, & 2015
Hewlett-Packard	S2	A	2001-2015	2014	2010	2009, 2012, & 2015
EMC Corporation	S3	A	2009-2015	2014	2011	2009, 2012, & 2015
Western Digital Corporation	S3	A	2011	2011	N/A	N/A
NetApp, Inc.	S3	A	2016	2016	N/A	N/A
IBM	S4	A	2002-2015	2014	2013	2009, 2012, & 2015
Xerox Corporation	S4	A	2009-2016	2015	2014	2009, 2012, & 2015
Computer Sciences Corporation	S4	A	2009-2016	2015	2009	2009, 2012, & 2015
Cognizant	S4	A	2014	2014	N/A	N/A

ISSN: 1941-6679-On-line Copy

eBay Inc.	S5	A	2012-2014	2014	2013	N/A
Cisco Systems, Inc.	S6	A	2005-2016	2015	2012	2009, 2012, & 2015
Qualcomm Incorporated	S6	A	2006-2015	2015	2014	2009, 2012, & 2015
Motorola Solutions, Inc.	S6	A	2014-2015	2014	2015	N/A
AT&T	S7	A	2006-2015	2015	2009	2009, 2012, & 2015
Verizon Communications	S7	A	2004-2015	2015	2009	2009, 2012, & 2015
Comcast	S7	A	2013	2013	N/A	N/A
DIRECTV	S7	A	2011-2014	2014	2013	N/A
CenturyLink, Inc.	S7	A	2014	2014	N/A	N/A
Time Warner Cable Inc.	S7	A	2012-2014	2012	2013	N/A
Intel	S8	A	2001-2015	2014	2015	2009, 2012, & 2015
Texas Instruments	S8	A	2010-2015	2014	2012	N/A
Applied Materials	S8	A	2007-2015	2014	2012	2009, 2012, & 2015
Broadcom	S8	A	2014	2014	N/A	N/A
SanDisk	S8	A	2013	2012 & 2013	N/A	N/A
Advanced Micro Devices	S8	A	2010-2015	2014 & 2015	2011	N/A
NCR Corporation	S2	O		Obtained in May, 2016	N/A	N/A
Micron Technology	S8	O		Obtained in	N/A	N/A

				May, 2016		
Jabil Circuit	S8	O		Obtained in May, 2016	N/A	N/A
Sanmina	S8	O		Obtained in May, 2016	N/A	N/A
Booz Allen Hamilton Holding Corp.	S4	O		Obtained in May, 2016	N/A	N/A
Amazon.com	S5	O		Obtained in May, 2016	N/A	N/A
Google	S5	O		Obtained in May, 2016	N/A	N/A
Facebook, Inc.	S5	O		Obtained in May, 2016	N/A	N/A
Corning Incorporated	S6	O		Obtained in May, 2016	N/A	N/A

*: S1-Computer Software; S2-Computer, Office Equipment; S3-Computer Peripherals; S4-Information Technology Services; S5-Internet Services and Retailing; S6-Network and Other Communications Equipment; S7-Telecommunications; S8-Semiconductors and Other Electronic Components;
**: A-Annual Report; O-Online Disclosure
***: DB-Dictionary Building

ISSN: 1941-6679-On-line Copy