Environment Systems and Decisions (2023) 43:416–432
https://doi.org/10.1007/s10669-023-09900-7

Data modeling and NLP-based scoring method to assess the relevance of environmental regulatory announcements

Heiko Thimm[1]

Accepted: 20 February 2023 / Published online: 21 March 2023

Abstract

The constantly growing body of global environmental legislation necessitates that corporate environmental compliance managers frequently assess the relevance of new regulations and regulation revisions for each of their sites. Companies are pressured to streamline and automate this crucial task through digital workflows and specialized IT-based assistance systems. This has recently piqued the interest of researchers working in different disciplines, such as intelligent systems, machine learning, and natural language processing. The article describes the latest results of our long-term research program on IT-based support for corporate compliance management, offering insights for these, and other disciplines. The context and the main aspects of environmental regulation announcements and the relevance assessment task are analyzed. An extensive conceptual data model is developed that serves as a foundation for tailoring a generic method to perform a relevance assessment that considers site-specific individual environmental compliance facts. The method uses heuristic data operations and various text processing techniques from the field of natural language understanding. In order to exemplify the method, two application scenarios are described in which the relevance of new waste management directives are assessed for a multi-site production company.

Keywords Corporate environmental compliance management · Heuristic relevance scoring · Text processing technology · Natural language processing · Legal artificial intelligence · Conceptual data modeling

1 Introduction

Announcements of new entities of environmental legislation, such as laws, acts, and directives, referred to in the following as 'environmental regulations' or just 'regulations' and announcements of revisions of already existing regulations must be continuously monitored. Among other corporate environmental compliance management (CECM) duties, this monitoring task is a central obligation for all business organizations. Whenever a new regulation or revision is announced, the relevance for the firm must be assessed. Any relevant regulation and its revisions need to be documented together with respective enforcement measures for auditing purposes and compliance checks. Companies are recommended to make use of a central regulation registry (Thimm 2015) to streamline operational compliance management tasks. The key data administered in a regulation registry are regulations, revisions, the results of relevance assessments of regulatory announcements, and measures that target enforcing compliance.

The accurate assessment of regulatory announcements and the provision of an up-to-date regulation registry can be perceived as a crucial requirement (Campbell Gemmell and Marian Scott 2013) that needs to be fulfilled to achieve full compliance with environmental legislation. Clearly, a firm's environmental compliance situation must be viewed as a continuous time-varying state (Thimm 2017b). Transitions from a (positive) compliance state into a noncompliance state where environmental legislation is violated may occur for many different reasons, including breakdowns of compliance enforcement measures, malfunctioning of infrastructure and equipment, human errors, organizational deficiencies, limited expertise, limited trust in the governmental compliance enforcement systems, sabotage acts, and environmental crime (White and Heckenberg 2012). Note in this context that any revision of product properties, production processes, and material logistic routines and involved

Heiko Thimm
heiko.thimm@hs-pforzheim.de

[1] School of Engineering, Pforzheim University, Tiefenbronnerstr. 65, 75175 Pforzheim, Germany

equipment need to be carefully evaluated in terms of potential compliance conflicts.

A recent research report of Good Jobs First, a nonprofit organization based in Washington, DC, (Mattera and Baggaley 2021) p. 6, describes the following: 'Over the past two decades, state regulatory agencies and attorneys have generally brought more than 50,000 successful enforcement actions against private sector entities for violations of clean air, clean water, and other environmental laws. Looking at cases with penalties of $5000 or more, the states have collected about $21 billion in fines, settlements, and other payments.' In the context of public debates of such numbers, industry associations often draw attention to the regulation density and the dynamic of environmental legislation, which has been growing dramatically during the last decades. It is argued that this trend, which is expected to become even stronger in future, has led to a highly complex and heterogeneous body of environmental legislation that imposes severe problems on the business world. In particular, multinational companies with many different production sites, supply chain partners, and customers around the world are already challenged today and even more in the near future by global environmental legislation, which is constantly being extended and revised by many different rule setters at various levels, including municipalities, counties, states, countries, and supra-national organizations.

Difficulties in finding people with suitable skills for the complex set of CECM tasks and budget constraints are reasons that some companies have outsourced or out-tasked the monitoring and relevance assessment of regulatory announcements. Typical contract partners in practice are law firms, environmental consultancies, and highly specialized software companies that offer curated environmental legislation content.

Regardless of whether the announcement monitoring and the relevance assessment is performed in house by corporate environmental compliance managers or completed by contractors, in general, IT-based assistance of these tasks may make it easier for firms to handle the challenges described above (Thimm 2017a). The research described in this article aims to investigate IT-based assistance for announcement monitoring and relevance assessments and to pioneer and test respective assistance tools. We do not target what is in legal informatics often referred to as 'legal machines' (Cyras and Lachmayer 2014), but our approach bears some commonalities to these works, particularly regarding the aspect of legal subsumption.

A generic data model is presented that combines both the modeling of domain knowledge and the modeling of the specific CECM context of firms. The data model is intended to serve as the foundation for a generic assistance system that computes relevance scores for new regulations. A second obvious building block of the targeted assistance system is a tailored scoring method that computes accurate relevance scores for the firm's sites that may need to comply with different sets of environmental regulations. This requires the method to carefully address the actual activity profile and the prevalent regulatory situation for each site. A first version of such a scoring method is proposed on the basis of the data model. The method combines intuitive analysis steps and analysis steps that apply standard text analysis techniques from the field of natural language processing (NLP). Two application scenarios are described in order to exemplify the principles of the method and to demonstrate the validity of it. In each scenario it is assumed that a new waste management directive has to be assessed for a multi-site production company. The sample data used to describe the assessment steps of the method are largely based on the real-world CECM data provided by an industry partner. The article proceeds as follows. Related work and an overview of the text analysis techniques considered in this work are described in Sect. 2 and Sect. 3, respectively. The main aspects of the regulatory announcements are investigated in Sect. 4. The data model is introduced in Sect. 5. An overview of the proposed scoring method is given in Sect. 6. Also, in Sect. 6 the application scenarios are briefly described, while corresponding detail data are given in the appendix. Concluding remarks are contained in Sect. 7.

2 Related work

Butler (2011) published theories to explain how green information systems can support organizational sense making, decision-making, and knowledge sharing. The work includes a conceptual model of the process of regulatory compliance gathering and the process of compliance decision-making. Several similarities between the Butler model and the concepts proposed in this research can be found. Butler looks at the processes as a whole abstracting from implementation aspects. In contrast, this research investigates specific decision tasks of these processes and describes a solution approach for implementation.

In an Irish case study (Butler and McGovern 2012), the company Napa Inc. was analyzed concerning major CECM issues. The researchers explored the fundamental compliance processes and challenges that firms face in their CECM practice in general, and in particular, with respect to the use of ICT for CECM tasks. The study is focused on product compliance, while this research targets compliance across all business functions, including manufacturing, logistics, supply chain management, and the interference of a firm's physical work space with the environment. Additionally, generic information system (IS) solutions for CECM are analyzed and a process-based conceptual model for CECM and a conceptual architecture of a CECM IS called the 'compliance knowledge management system' are described. Similar aspects of CECM research have been addressed by a research group at Pforzheim University (Thimm 2015). The group proposed a comprehensive process model and an information system approach for CECM. The more recent work of the same group discussed a conceptual framework for cloud-based assistance of CECM practitioners (Thimm 2018).

A reference model for an environmental management information system for compliance management that makes comprehensive use of business intelligence concepts has been proposed by Freundlieb and Teuteberg (2009). Kerrigan (2003) of Stanford University proposed a software infrastructure that offered assistance for CECM tasks based on semantic technologies. A research group of IBM developed an approach for compliance automation through the use of event monitoring rules (Giblinet al. 2006). Wizards for conveying environmental information and helping people complete environmental management tasks, for example, are described in (Braun et al. 2004).

We also searched for related work in the research literature on business process management and decision modeling. The results of these search efforts suggest that the CECM field has not been at the focus of business process management research thus far. Most articles address corporate sustainability management at a strategic level and thus, focus on higher-level perspectives, such as the business case level (Schaltegger et al. 2012) or the business model level (Geissdoerfer et al. 2018). It appears that work on decision modeling for the domain of corporate sustainability management has largely not addressed issues of the CECM field.

In recent research studies advanced NLP methods and Machine Learning have been applied to extract specific knowledge items from text documents of the construction domain, such as contractual risk clauses (Moon et al. 2022), requirements (Hassan and Le, 2020), and legal and contractual matters (Hassan et al. 2021). The results of these studies may offer promising research avenues for the computer-based relevance assessment of environmental regulations targeted in this work. However, one needs to consider that several fundamental differences exists between the text documents of the two domains. Therefore, it cannot be expected that the extraction methods for construction documents will provide reasonable results for documents of the environmental compliance management domain when just the used ontology or dictionary is tailored to the new domain.

NLP and AI in general have already been applied in the legal domain for several decades (Dale, 2019). However, today's large interest of researchers and practitioners in what is known as LegalTech was mainly caused by recent advancements in AI. According to Haney (Haney 2019), p. 3, 'Today, NLP is the most commonly used method of AI in the practice of law.' In a recent journal article, Dale (Dale, 2019) defines the following five areas of legal activity where NLP is playing an increasing role: legal research, electronic discovery, contract review, document automation, and legal advice. The particular CECM tasks that this research targets to support based on NLP belong most likely to the area of legal research characterized by 'finding information relevant to a legal decision' and 'electronic discovery' characterized by 'determining the relevance of documents to an information request.' However, only a few works can be found where the application of different NLP methods in these fields were studied.

3 The text analysis methods considered

Today, there exist a broad variety of different text analysis methods (Anandarajan et al. 2019; Bird et al. 2009; Gudivada and Rao, 2018). In recent years, the traditional methods mostly developed by the NLP and the Information Retrieval research community have been complemented by new methods that address what is often characterized as 'Big Data' through the use of machine learning approaches, especially the use of deep learning (Ghavami 2020). For the proposed relevance scoring method, four traditional NLP text analysis methods have been chosen that partially build on each other. In the following, a brief overview of these methods is given.

Keyword Frequency Analysis (KFA). As the name suggests, this method computes the frequencies of words or phrases as they appear in a text. The method often serves as the basic building block of higher-level text analysis methods (Illinois University Library 2022). However, for some use cases, the raw word counts or percentage numbers for words may already reveal useful insights.

Named Entity Recognition (NER). It is the objective of this method to recognize and extract specific types of entities, such as names of people, organizations, machine elements, locations, times, quantities, monetary values, percentages, and more in a text (Foleyet al. 2018; Jagota

2020). Typical use cases of the method locate entities to pull specific information from a text, to discover the subject of a given text, and to discover relationships among the entities. The method has also been used to improve web searches and document indexing and to support building an ontology. In these and other use cases, an NER analysis serves as a first preprocessing step, which is followed by other specialized text analysis processing steps.

Different NER algorithms have been developed. Dictionary-based algorithms use dictionaries of values of every entity type that is to be recognized. One of the main drawbacks of dictionary-based approaches is that they cannot effectively handle ambiguity. Algorithms that use probabilistic dictionaries particularly address the ambiguity of words. Pattern-based algorithms use regular expressions. These algorithms are most applicable when the targeted entities are best described by structural patterns. Several further traditional non-dictionary-based NER algorithms exist, such as rule-based algorithms and machine learning-based NER algorithms, which require large, annotated training data. In comparison to the simple text scanning and finding hits of the classical dictionary-based algorithms, these algorithms are usually far more complex. In some NER applications, a combination of multiple approaches has been used (Keretna, et al. 2014).

Document Subject Identification (DSI). Various techniques to identify the main subject(s) of a single document are described in the literature. According to D'Hondt, these techniques can be divided into two categories (D'hondt et al. 2011) p. 3784: '… techniques using statistical information extraction techniques and those exploiting lexical cohesion.' For both categories, algorithms have been proposed that are based on machine learning approaches. The 'latent Dirichlet allocation' (LDA) algorithm is a frequently used algorithm that explores subject probabilities from available statistical data.

In this work, we focus on DSI methods that exploit lexical cohesion without the use of machine learning techniques through a combination of a dictionary-based NER analysis and statistical methods, such as cluster analysis. It has been argued that the results of dictionary-based DSI algorithms are dependent on the semantic resources available for a specific text; therefore, the setup is limited to the text (D'hondt et al. 2011). However, this drawback can be at least partially overcome through the use of a comprehensive and well-developed dictionary. A methodology for building dictionaries was proposed by a research group at Carleton University (Denget al. 2019). This methodology suggested obtaining an initial version of the intended dictionary from an existing context-specific text corpus by applying an NER analysis.

Document Similarity Analysis (DSA). The goal of DSA is to measure the pairwise similarity between the text documents (Elia 2020). Corresponding techniques can be divided into DSA methods for this task that work on a lexical level (i.e., surface closeness of two text instances), meaning that they use only the words in the sentence, and methods that go beyond that and measure semantic similarity (i.e., similarity of meaning). Methods to measure semantic similarity attempt to explore the actual meaning behind words or the entire phrase in context. Clearly, this is a far more difficult measuring task than the task to measure lexical similarity. At the current stage of this research, it is focused on the use of similarity scores that just measures the lexical similarity between documents. One of the earliest techniques to compute such scores is the vector space model (Shajalal and Aono 2019). Methods that are built on this model compute similarity scores in two steps. First, the documents are transformed into a vector representation and then a similarity score is computed using a vector distance calculation formula. The 'Term Frequency-Inverse Document Frequency' (TF-IDF) vectors are frequently used in the vectorization step of many implementations of this method (Neto 2021). Often, for the similarity score calculation, the methods either use the cosine similarity, which has a value ranging from − 1 to 1, or the Euclidean distance. Other distance metrics are Jaccard, Manhattan, and Minkowski.

4 Environmental regulatory announcements

Today's growing body of environmental legislation consists of laws, acts, ordinances, statutory commands, treaties, subordinances, and other forms of environmental obligations for the business world (Campbell Gemmell and Marian Scott, 2013; German Environment Agency 2019; Ruhl 1997) that we subsume in the following by the notion of an 'environmental regulation' or just a 'regulation.' One of the main drivers of today's strong environmental regulation dynamics is the climate change action plan of the United Nations Organization (UNO), which encourages politicians around the world to tighten environmental laws.

In general, the empowerment of environmental authorities to issue regulations is usually limited to a particular territory. Examples of territories of authorities given in a hierarchical order are a city, a county, a state, a country, and the territory of a supra-national union of countries, such as the territory of the European Union.

When a decision for a new regulation or for a revision of an already established regulation has been made, usually the greater public is informed by the authority through a regulatory announcement. Typically, the announcements are text documents that contain a copy of the authority's original regulation text or of the revision text. Additionally, announcement documents may contain metadata about the authority, metadata and general data about the regulation, and context-specific background information. The announcement documents are published through various channels, such as the internet, print media, special governmental media, online databases, special information agencies, and service providers. Some special service providers have the provision of a curated database of announcement documents as their business model along with explanations, guidelines, and recommendations for corporate compliance managers.

Clearly, announcements should be made in a timely fashion to give firms enough time to complete checks concerning the relevance for the firm and, when needed, to react through respective compliance enforcement measures. As of today, there does not exist a common format, nomenclature, or common language style for regulatory announcements. Abstract sentences with many technical terms and references to entities of the current body of environmental legislation are frequently used to describe new regulations and revisions of existing regulations. Therefore, much experience and effort are required to obtain the criteria that are to be checked to know if a regulation applies to a firm. The task of obtaining this knowledge through a corresponding investigation bears some similarities to what lawyers refer to as 'legal characterization' or 'legal subsumption.' A description of the theoretical foundation of the notion of subsumption, for example, is available in (Cyras and Lachmayer, 2014).

When regulatory announcements are explored for a relevance assessment, those items of the CECM work field that are addressed by the regulation and are prevalent in the business practice of the firm need to be investigated. We refer to these investigation items by 'items of CECM concern' (Thimm 2022). Table 1 contains several simple examples for this concept. It can be expected that in the CECM practice of typical multi-site production companies, many of the items of CECM concern account for the firms' Scope 1 and Scope 2 greenhouse gas emissions of the Greenhouse Gas Protocol (WBCSD and WRI, 2004).

To explore a new regulation or revision, we ask the following questions: What general field(s) of environmental law is (are) being addressed? What is the spatial/geographical boundary of the regulation? What particular business aspects (e.g., product properties, aspects of the production method, material usage) are being addressed? Which types of items of CECM concern are the focus of the regulation? What (pollution) limits (e.g., waste water temperature) are addressed? What conditions for exceptions are described? Which other entities of environmental legislation are related to the regulation and should be considered in the investigation? The answers to each of the questions need to be put into the context of the firm. The particular company situation with respect to contextualized versions of the above questions needs to be explored. Additionally, sample questions to ask include the following: Does the firm fit to the particular spatial/geographical scope of the regulation? Do the business activities of the company include particular entities addressed by the regulation? Do the particular entities satisfy the specific constraints set for the type of entities?

A relevant regulatory announcement may require compliance enforcement measures that are targeted at the particular business aspect and type of item of CECM concern explored in the relevance assessment. Table 1 describes some simple examples of measures and measure categories that may be taken into consideration for particular types of items of CECM concern.

Table 1 Examples of business aspects and concepts to be addressed in CECM tasks

Business aspect	Type of item of CECM concern	CECM measure	Measure category
Factory staff	Skill level	Instruction, training	Training/education
Raw material pipeline		Installation of leakage detection device	Implementation of special compliance enforcement equipment
Manufactured product	Hazardous Chemical substances	Replacement of hazardous substances	Product revision
Production system	Waste air	Installation of emission filter technology	Infrastructure revision
Production processes	Concrete waste	Collection of sorted waste in special waste thins	Implementation of special compliance enforcement equipment
Production processes	Concrete waste	Installation of sign boards with waste handling instructions	Information measure
Production processes	Waste water	Installation of waste water treatment facility	Infrastructure revision

Note that regulatory announcements require firms to perform relevance assessments of the environmental legislation. Changes in the firm's business model, business processes, production processes, product portfolio, growth activities, and other movements to a new status quo may affect business aspects that are relevant for the work field of CECM. Firms are further obligated to assess these changes in terms of their conformance with the relevant environmental legislation.

Additionally, another relevant fact of this research is that companies are expected to maintain a (digital) regulation registry also known as regulation cadaster (Thimm 2015). The registry has to store all regulations and the respective relevance assessment results together with potential measures taken to enforce compliance. Clearly, a regulation registry (especially in digital form) is one of the most essential tools for effective environmental compliance management and other tasks of corporate environmental management, such as audit and inspection management and permit management. Additionally, annual environmental and sustainability reports, in addition to other information, are usually composed of data stored in the regulation registry, such as the number of environmental incidents, compliance violations (Thimm 2019), measures, and savings obtained by the measures.

5 The data model

In general, conceptual data modeling (Robinsonet al. 2015) is a discipline where a particular application domain or 'universe of discourse' is modeled, for example, as a founding step of a database development project. Conceptual models based on the well-known Entity Relationship Modeling (ERM) method address two main concepts at the intentional level: entity types depicted in Entity Relationship Diagrams (ERD) as labeled boxes and relationship types depicted as labeled rhombuses. The properties of both concepts are also addressed in the ERM method and in ERD diagrams

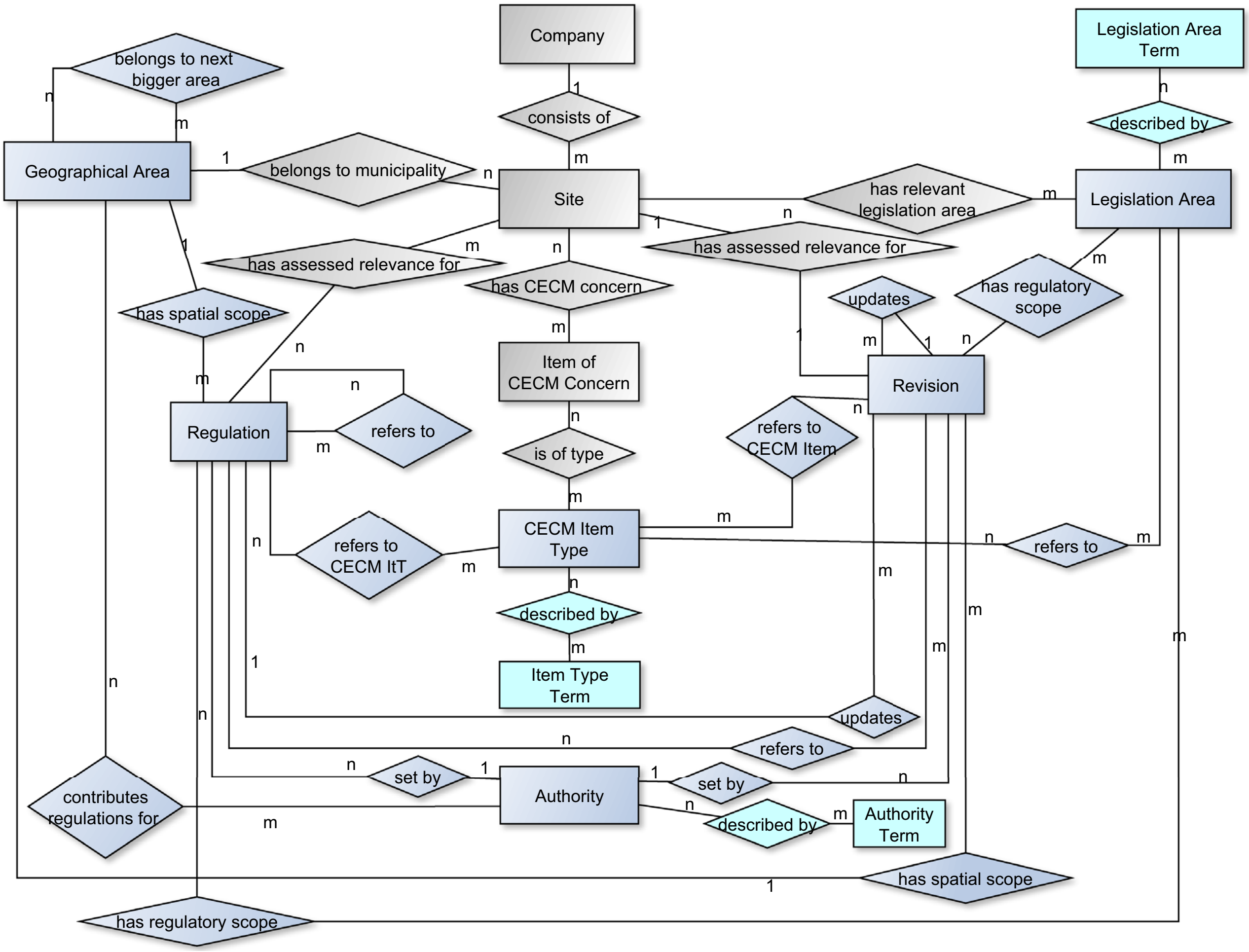


Fig. 1 Conceptual data model of the CECM work field with a particular focus on the relevance assessment task (overview version without properties)

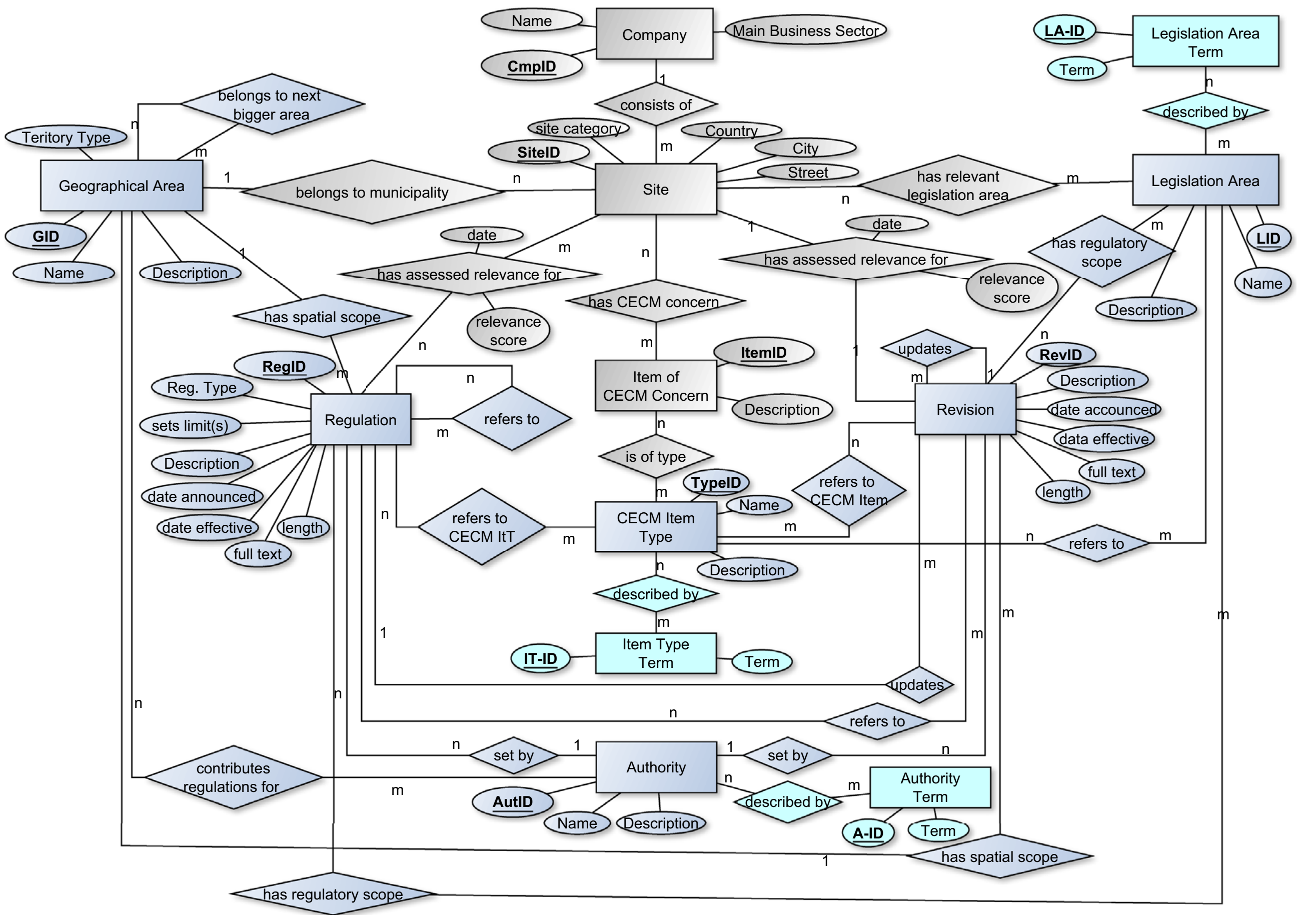


Fig. 2 Refined version of conceptual data model with properties

displayed as labeled circles. In ERD diagrams the maximum number of relationship instances in which entities can participate are indicated through respective cardinality numbers that are associated with the corresponding edges.

For the domain of CECM, the researchers devised a conceptual data model using the ERM method with cardinality information given in the classical Chen notation (Chen 1976). In this notation, n and m stand for two distinct numbers with arbitrary positive integer values. The resulting ERD in Fig. 1 is displayed in a version that abstracts from the properties in order to give a first overview of the complex data model. The full version of the model that includes the properties is contained in Fig. 2. All major concepts, issues, and definitions described in the previous section with respect to the relevance assessment CECM tasks are addressed in the model. The gray parts of the ERD are abstractions for company-specific aspects that are to be considered in the relevance assessment task. The ERD elements in blue address the domain knowledge required for this CECM task. Among these elements are the entity types (depicted in light blue color): ‘Legislation Area Term,’ ‘Item Type Term,’ and ‘Authority Term,’ which model the terms of three dictionaries. As described in the next section, these three dictionaries serve the NLP-processing steps of the proposed relevance assessment method.

The company-specific model part consists of the three entity types labeled ‘Company,’ ‘Site,’ and ‘Item of CECM Concern.’ Clearly, ‘Company’ models a company and ‘Site’ models an individual site of a company. The headquarters of a company (i.e., main place where it is registered) is modeled through the attribute ‘site category.’ Companies can consist of multiple sites that are modeled by respective cardinalities of the relationship type, ‘consists of.’ The relationship type ‘belongs to municipality,’ in general, models the sites that are associated with governmental structures. The model focusses through the relationship type on the smallest relevant ‘governmental administration unit,’ which usually is the municipality (German Environment Agency 2019). Note that the containment of municipalities to larger 'governmental administration units' (e.g., states, countries, supra-national organizations) is also addressed in the CECM knowledge part of the model. The relationship type, 'has relevant legislation area,' models some areas of environmental legislation that are relevant for the site, while others are not. Moreover, whether the areas of environmental legislation are relevant for a site largely depends on what is going on at the site (e.g., production site, storage sites, development sites, administration) (German Environment Agency 2019). The two relationship types with the identical name, 'have assessed relevance for,' model the site-specific relevance of regulations and revisions.

Of the modeled CECM domain knowledge, the entity type, 'Geographical Area,' models geographical territories relevant to environmental legislation. In addition, some territories may be contained in other larger territories, which are addressed by the unary relationship type, 'belongs to next larger area.' 'Legislation Area' models the regulation areas that are considered areas of environmental legislation, such as 'water protection,' 'waste,' 'chemicals,' and 'air pollution' (German Environment Agency 2019). Dictionary terms addressed by the entity type, 'Legislation Area Term,' are, for example, 'waste water,' 'effluent,' 'sewage,' 'lead concentration,' 'particle content,' and 'leakage.' 'CECM Item Type' models aspects through which companies interfere with the environment, such as waste water, waste air, concrete waste, air pollution, and resource consumption (Thimm 2022). Also, the same type is used to model aspects that impose risks for the environment, such as explosive composites, hazardous material, leakages, water drain, and storm weather conditions. Respective dictionary terms modeled by the entity type 'Item Type Term' are terms frequently used for the contextualization of regulations and the detailed specification of limitations, threshold values, and critical values. Sample terms that refer to waste water are 'natural river discharge,' 'temperature limit,' and 'monitoring obligation.' Sample terms that refer to hazardous chemical substances are 'arsenic,' 'lead,' 'benzene,' 'chromium,' and 'toluene.' Clearly, the entity type, 'Authority,' models authorities that are entitled to issue environmental regulations and revisions (German Environment Agency 2019). Dictionary terms used in regulatory announcements to refer to these authorities are addressed by the entity type, 'Authority Term.' Examples of such terms include 'Environmental Protection Agency,' 'EPA,' 'European Environmental Agency,' 'EEA,' 'German Federal Environmental Agency,' 'UBA,' and 'Istituto Superiore per la Protezione e la Ricerca Ambientale,' 'ISPRA.' In general, these authorities issue regulations and revisions of regulations that are addressed in the model through the entity types, 'Regulation' and 'Revision.' Furthermore, some regulations set limits, for example, on exposure levels and content shares, and addressed by the attribute 'sets limit(s)' of the type 'Regulation.' The attribute, 'announcement text,' modeled for type, 'Regulation' and type 'Revision,' serves as a proxy for the respective announcement document. The binary relationship type, 'updates,' models the revisions that update existing regulations. The relationship type, 'refers to,' stipulates that a revision may refer to multiple regulations and vice versa and a regulation may refer to multiple revisions. The revisions that may lead to changes of earlier revisions of regulations are modeled by the unary relationship type 'updates.' Both regulations and revisions refer to items of the CECM concern expressed by the two identical relationship types, 'refers to CECM Item.' The geographical scope of regulations and revisions is modeled by the relationship type, 'has spatial scope.' The regulatory scope is modeled by the relationship type, 'has regulatory scope.' The regulations may refer to other regulations of the environmental legislation, which are modeled through the unary relationship type 'refers to.'

6 Toward an NLP-based relevance assessment method

On the basis of the conceptual data model described above, the researchers developed a first version of a relevance assessment method that builds on the possibilities of today's text processing technology. In particular, the NLP methods described in Sect. 3 are applied. The proposed assessment method computes site-specific numeric scores for a (new) regulation. A score indicates to what extent the regulation is relevant for the specific site. In principle, through exactly the same steps used by the method, relevance scores for revisions of regulations can also be computed. The relevance scores are obtained from heuristic rules that are derived from both intuition and observations of CECM practitioners. The rules are implemented in a multistep scoring scheme that investigates the content of regulations and CECM data at the firm level and at the firm site level. The data are contained in a central repository that is based on the conceptional data model described in Sect. 5 and in the following referred by the 'Environmental Compliance Knowledge and Data Repository' or in short 'EKDR.'

Table 2 Variables of proposed relevance assessment method

Variable	Description
R_{new}	A new regulation document
$S = \{S_1, S_2, \ldots, S_k\}$	A set of i: = 1, …, k company sites S_i
$R_i = \{R_{i,1}, R_{i,2}, \ldots, R_{i,g}\}$	A set of j: = 1, …, g regulations $R_{i,j}$ already assessed for company site S_i
$as_i \in [0; 2]$	A relevance assessment score of company site S_i concerning a particular R_{new}
$xs_{i,q} \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$	A text similarity score computed by function DSA-REG which measures the pairwise lexical similarity between the textual content of two regulation documents; score $xs_{i,q}$ is specific to site S_i
$X_i = \{xs_{i,1}, xs_{i,2}, \ldots, xs_{i,v}\}$	A set of q: = 1, …, v text similarity scores $xs_{i,q}$ obtained for company site S_i
$ss_i \in [0; 1]$	A similarity score of company site S_i
$cs_i \in [0; 1]$	A normalized coverage score of company site S_i
$c\hat{s}_i \in [0; 4]$	A calculative coverage score of company site S_i
$rs_i \in [0; 2]$	A relevance score of company site S_i
$hsR_i \subset R_i$, $lsR_i \subset R_i$	Two disjunct sets of regulations already assessed for site S_i with set hsR_i containing regulations with a high similarity to R_{new} and the set lsR_i containing regulations with a low similarity to R_{new}
hts_i, $lts_i \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$	text similarity scores with hts_i obtained from the high scores contained in set hsR_i and lts_i obtained from the set of low scores contained in set lsR_i
$T_i = \{T_{i,1}, T_{i,2}, \ldots, T_{i,k}\}$	A set of d: = 1, …, k technical terms $T_{i,d}$ that describe the regulatory context of a particular site S_i for a specific regulation area
$TC_i = \{TC_{i,1}, TC_{i,2}, \ldots, TC_{i,h}\}$	A particular set of c: = 1, …, h technical terms $TC_{i,c}$ obtained by the function TCA-CECM
$w \in \mathbb{N}$	Word count value of a particular regulation document obtained by function WCT
$wq1_i$, $wq2_i$, $wq3_i \in \mathbb{N}$	Values of the three lower quartiles computed by the function IQR-WCO for the word count frequencies of the set of regulation documents that are already assessed for site S_i
$s_i \in \{0.25, 0.5, 0.75\}$	Scaling factor used to calculate a coverage score relative to both the length of the new regulation R_{new} and the length of the regulations already assessed for site S_i

6.1 Scheme of method and a 'Toy Example'

The proposed scheme performs a relevance assessment for a new regulation document in several steps that are grouped into two subsequent phases, an initial scoping phase that is followed by a scoring phase. The variables used to describe the scheme are defined in Table 2.

The scoping phase obtains the geographical and regulatory scope of the regulation in an initial analysis step. Then, the firm sites that are within the scope of the new regulation are explored and further investigated in the scoring phase. In the scoring phase, the CECM-specific context of the site is explored. This includes an analysis of the already registered regulations and of the registered items of CECM concern. An aggregated numeric relevance assessment score $rs_i \in [0; 2]$ is obtained for every site. Consequently, the interval [0; 2] serves as scoring scale A score of 0 means that the regulation is not relevant at all, whereas a maximum score of 2 indicates that the regulation is definitely of utmost relevance for the site. Scores above 0.8 are considered to indicate regulations to be of relevance.

The following description of the two phases is focused on the general high-level algorithm, the integrated NLP methods, and the data of the EKDR repository used by the processing steps. The data analyses are performed by functions that are defined in the following list. The document with the particular new regulation is passed to each function through parameter < regdoc > . Note that the given informal definitions abstract from the common NLP preprocessing steps, such as lemmatization and stemming (Anandarajan et al. 2019).

- A function denoted WCT(< regdoc >) is defined that computes the word count of a regulation document denoted by < regdoc > .
- A function denoted by DSI-AUTHORITY(< regdoc > , < authority_dict >) is defined that identifies the authority that issued the new regulation. The function performs a dictionary-based NER analysis using a term dictionary denoted by < authority_dict > . The terms identify common authorities that act as environmental rule setters. When several authorities are extracted, sta-

tistical operations are used to identify the correct issuer of the regulation.

- A function denoted by DSI-LEG-AREA(< regdoc, < legarea_dict >) is defined that identifies the particular environmental legislation area addressed by the new regulation. Using a dictionary that contains terms to identify regulation areas denoted by < legarea_dict >, an NER analysis is performed. The final selection from several extracted regulation areas is made through statistical operations.
- A function denoted by DSA-REG(< regdoc >, < scored_regdoc >) is defined that computes a text similarity score $xs_{i,q} \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. The text similarity score measures the pairwise lexical similarity between the document < regdoc > and a relevant already scored document denoted by < scored_regdoc >. The function compares the documents and based on a linear value assignment model returns the lowest score value of 0 when no similarity is found. Otherwise, a larger score is returned with the score value being derived from the extent of similarity. That is, a score value of 10 indicates maximum similarity.
- A function denoted TCA-CECM(< regdoc >, < cecm_dict >) is defined that searches the document < regdoc > for terms contained in the term dictionary denoted by < cecm_dict >. The function results in the set $TC_i = \{TC_{i,1}, TC_{i,2}, \ldots, TC_{i,w}\}$ that contains the w terms found in the document.
- A function denoted IQR-WCO(< regcollection >) is defined that determines the frequency distribution of the word counts of a set of regulation documents denoted by < regcollection >. The function results the values of the lower quartile, the middle quartile (i.e., the median), and the upper quartile denoted by $wq1, wq2, wq3, \in \mathbb{N}$. Following general definitions of the statistics discipline, the value of wq1 implies that 25% of the set of documents are having a word count lower or equal than wq1. Likewise, 50% of the documents are having a word count lower or equal than wq2, and 75% of the documents are having a word count lower or equal than wq3.

Scoping phase. The scoping phase is performed through the following four steps:

(1) Obtain the word count w of the regulation R_{new} through function WCT(< regdoc >).

(2) Check the new regulation R_{new} for common metadata patterns used to specify the authority. When the document does not contain proper metadata, then perform function DSI-AUTHORITY(< regdoc >). Retrieve from the EKDR repository the geographical area for which the identified authority sets regulations. Proceed with the retrieved geographical area being used as the geographical scope of the regulation.

(3) Check the document for metadata patterns used to specify the regulation area. When no proper metadata are found, then perform function DSI-LEG-AREA(< regdoc >). Continue with the recognized regulation area being used as the regulatory scope of the regulation.

(4) Retrieve the particular set of company sites $S = \{S_1, S_2, \ldots, S_k\}$ from the EKDR repository where each site S_i $(i = 1, \ldots, k)$ is within the geographical and the regulatory scope of the regulation.

Scoring phase. In this phase, the k sites of the set of sites S are assigned relevance assessment scores $as_i \in [0; 2]$. The scores are obtained through the formula, $as_i = ss_i + cs_i$. The component score $ss_i \in [0; 1]$, referred by similarity score, is obtained through a similarity analysis. The component score $cs_i \in [0; 1]$, referred by coverage score, is obtained through a coverage analysis. The intuitions behind these analyses and their principle processing steps to determine the scores are described in the following.

Similarity analysis. The rationale behind the similarity analysis is that when many similar regulations can be found that are relevant for the site, then the new regulation R_{new} is also likely to be relevant. On the basis of this rationale, an algorithm has been devised that consists of the following five steps that result a similarity score ss_i for a particular site S_i with respect to R_{new}:

(1) Of the regulations already assessed for site $S_{i,}$ obtain from the EKDR repository the particular set of regulations that have the same regulatory scope as R_{new}. Filter out for this set the set of $j := 1, \ldots, g$ regulations $R_i = \{R_{i,1}, R_{i,2}, \ldots, R_{i,g}\}$, which have been assessed to be relevant for site S_i.

(2) Perform a pairwise comparison between each regulation of the set R_i and R_{new} through function DSA-REG(< regdoc >, < scored_regdoc >) to obtain a corresponding set of $q := 1, \ldots, v$ text similarity scores $X_i = \{xs_{i,1}, xs_{i,1}, \ldots, xs_{i,v}\}$.

(3) Use the text similarity scores of set X_i to partition the set R_i in two subsets:

- a subset $hsR_i \subset R_i$ that contains the $j = 1, 2, \ldots, m$ (with $m < = v$) regulations $R_{i,j}$ for which a relative high similarity to R_{new} was determined such that $xs_{i,j} \geq 4$ for all $R_{i,j} \in hsR_i$
- a subset $lsR_i \subset R_i$ that contains the $p = 1, 2, \ldots, n$ (with $n < = v$ and $n + m = v$) regulations $R_{i,p}$ for which a relative low similarity or no similarity to R_{new} was determined such that is $xs_{i,p} < 4$ for all $R_{i,p} \in lsR_i$

(4) From set hsR_i obtain a specific score referred by high text similarity score or just high score denoted by $hts_i \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. When set $hsR_i = \{\}$ then consider a high score of $hts_i = 0$. Otherwise, assign hts_i the rounded mean score of the elements of set hsR_i. Likewise, obtain from set lsR_i a specific score referred by low text similarity score or just low score denoted by $lts_i \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Apply the rule used to obtain the high score in order to obtain the low score lts_i from the set lsR_i.

(5) For R_{new} compute the site-specific similarity score $ss_i \in [0; 1]$ from the high score hts_i and the low score lts_i through the following formula:

$$ss_i = \left(hts_i - lts_i\right)/10 \ldots \text{ iff}\left(hts_i - lts_i\right) > 0 \text{ otherwise } ss_i = 0.$$

Coverage analysis. Recall from above that the EKDR dictionary contains the regulatory context of a site S_i which is given by the CECM items that are associated with S_i. The coverage analysis focuses exactly at these items. The rationale behind the analysis is that when a relatively large number of a site's CECM items is contained in the description of the new regulation, then the new regulation is likely to be relevant. The total word count of the new regulation serves as the point of reference to obtain an appropriate relative measure concerning the regulation's number of CECM items. Using this rationale as the guiding principle, an algorithm of five steps has been developed. It computes a coverage score cs_i for a new regulation R_{new} and a particular site S_i as follows:

(1) Obtain from the EKDR dictionary the set of $d = 1, \ldots, k$ technical terms denoted by $T_i = \{T_{i,1}, T_{i,2}, \ldots, T_{i,k}\}$ that describe the regulatory context of S_i regarding the particular regulation area of R_{new}.

(2) Perform function TCA-CECM(<regdoc>, <cecm_dict>) with the regulation R_{new} and the set T_i being used as actual parameters. The function results in the set of $c = 1, \ldots, h$ terms $TC_i = \{TC_{i,1}, TC_{i,2}, \ldots, TC_{i,h}\}$ with each of the h terms being contained in both, in the set T_i and the regulation R_{new}.

(3) Perform function IQR-WCO(<regcollection>) with the set of regulations R_i obtained in the first step of the similarity analysis being used as actual parameter. This results the interquartile values $wq1_i$, $wq2_i$, and $wq3_i$ of the word frequency distribution of R_i. Apply the following rule to obtain from the result a scaling factor $s_i \in \{0.25, 0.5, 0.75\}$ that indicates how the word count w of R_{new} compares to the word count frequency of R_i:

$$s_i = 0.25 \text{ iff } w < wq1_i$$
$$s_i = 0.5 \text{ iff } wq1_i \le w \le wq2_i$$
$$s_i = 0.75 \text{ iff } wq2_i \le w \le wq3_i$$
$$s_i = 1 \text{ iff } w \ge wq3_i$$

According to this rule, a small scaling factor s_i is used when many regulations of R_i exceed the word count w of $R_{new.}$ Conversely, a large scaling factor s_i is used when many regulations of R_i have a word count that is lower than w.

(4) For the given site $S_{i,}$ obtain a coverage score $cs_i \in [0; 1]$ that is specific to regulation R_{new} using the cardinality of the term set TC_i denoted by $|TC_i|$ and the cardinality of term set T_i denoted by $| T_i |$. When $| T_i | = 0$ then use a coverage score $cs_i = 0$. When $| T_i | \neq 0$ then obtain cs_i in two steps. First, obtain a calculative coverage score $c\hat{s}_i \in [0; 4]$ through the following formula that uses the scaling factor s_i:

$$c\hat{s}_i = \left|TC_i\right|/\left(\left|T_i\right| \cdot s_i\right) \text{ with } \left|T_i\right| \neq 0.$$

In a second step apply the following transformation to obtain the targeted (normalized) coverage score cs_i from the calculative coverage score $c\hat{s}_i$: $cs_i = c\hat{s}_{i\ \ldots\ldots}$ iff $c\hat{s}_i \le 1$, otherwise $cs_i = 1$.

Practical 'Toy Example' to demonstrate the principles of the method. In the following, the scheme of the method described above is exemplified through two fictive application scenarios. The scenarios are invented from experience and also from some real data of an industry partner which is a German production company with multiple sites in Europe. The details of each scenario are described in the tables contained in the appendix.

A German manufacturing company is assumed which at three sites in Germany (Heilbronn S1, Karlsruhe S2, Bochum S3), at one site in Italy (Turin S4), and at one site in Austria (Orth S5) produces a set of diverse semi-finished goods for the furniture industry and the household appliances industry. Over the years a number of 69 regulations and revisions, respectively, concerning the area of waste management have already been assessed by the CECM specialists for each of the German sites. About the same number of regulations/revisions for waste management have also been assessed for the other two sites. The scenarios focus on environmental legislation for waste management and assume that a new regulation is set by the German Environmental Agency ('Umweltbundesamt'). The first scenario exemplifies the major steps of the scheme in order to obtain relevance assessments scores for a new waste management regulation concerning handling of end-of-life wood. In the second scenario, the assessment steps are performed for a new regulation concerning handling of end-of-live vehicles.

Note that the numbers used for the scenarios are obtained from respective real-world regulations and investigations of the regulation registry of the industry partner. In both scenarios in the initial scoping phase, the regulatory scope of the regulation is explored and the set of sites to be further investigated through the scheme's scoring phase is narrowed down to the three German sites S1, S2, and S3. Below, for each scenario the major results of the scoring phase and the final result are described for these three sites.

Assessment of regulation concerning handling of end-of-life wood. For the two production sites S1 and S3 that produce semi-finished goods for the furniture industry similarity scores of $ss1 = 0.6$ and $ss3 = 0.7$ are obtained. These scores reflect the fact that due to their focus site S1 and site S3 naturally deal with wood and therefore the method obtains a relatively large number of relevant regulations similar to the regulation for handling end-of-life wood. The relative low score of $ss2 = 0.1$ obtained for site S2 results from the site's specific production focus on goods for household appliances which logically implies a lower number of relevant similar regulations ($|R2| = 8$). That the new regulation is of much higher relevance for the sites S1 and S3 than it is for site S2 is also indicated by the respective coverage scores $cs1 = 0.96$ and $cs3 = 1$. These scores result from a comparison of the terms that describe the CECM-specific aspects of the sites with the content of the regulation document. The resulting aggregated relevance assessment scores $rs1 = 1.56$ and $rs3 = 1.7$ indicate that the new regulation on handling of end-of-life wood is of relevance for site S1 and site S3 but not of relevance for site S2, S4, and S5.

Assessment of regulation concerning handling of end-of-life vehicles. Obviously, the regulation on handling end-of-life vehicles addresses matters that are largely not of relevance for the sites S1, S2, and S3. Hence, the similarity scores $ss1 = 0.1$, $ss2 = 0$, and $ss2 = 0.2$ of the three sites are close to the minimal score of 0. The same holds true for the sites' coverage scores $cs1 = 0.11$, $cs2 = 0$, and $cs3 = 0.29$. Consequently, the regulation is not of relevance for any of the three sites which is indicated by the respective relevance assessments scores, $rs1 = 0.21$, $rs2 = 0$, and $rs3 = 0.49$.

6.2 Domain knowledge and company data: acquisition and curation

The proposed assessment method computes relevance scores based on domain knowledge and company-specific data. In various processing steps, the required items are retrieved from the ECKD data repository, which is derived from the data model described in Sect. 5. In the following, the major challenges and issues of the acquisition, the population, and the maintenance of these items of domain knowledge and company-specific data are discussed.

Dictionaries for the text analyses. The relevance assessment method builds on three-term dictionaries that are used to extract information from a new regulation document. The dictionaries < authority_dict > and < legarea_dict > are used to extract the authority and the regulation area, respectively. The dictionary < cecm_dict > is used to explore items of CECM concern. In general, there are a number of alternatives to obtain initial versions of the term collections. First, NER analyses can be used to compute the term collections from a suitable document subset of an environmental law text corpus, such as the EUR-Lex text collection of documents about the European Union law (European Union 2022). A second alternative is to use a suitable subset of the terms specified in environmental reporting standards, such as GRI (GRI 2022) or CDP (CDP 2022). The third alternative is to select a proper subset of the topics defined by the EUROVOC topic hierarchy, which contains almost 4000 categories concerning different aspects of European law (Filtz et al. 2019). The initial term collections obtained through one of these alternatives or a combination of the alternatives are to be tested and optimized by CECM experts during trial runs of the assistance system. To keep the dictionaries up-to-date and accurate, from time to time, their content also needs to be curated by the CECM experts.

Company-specific data. Clearly, the method's potential capability to compute company-specific relevance scores builds on processible data that specify the CECM context of the firm. Acquiring knowledge about all relevant data items and specifying these data items in a corresponding repository may seem to require substantial efforts, especially for large companies with many sites. However, much of the data may already be available in existing systems, such as corporate environmental management information systems, environmental, health and safety (EHS) systems, regulation cadasters, and other systems of the corporate information system landscape, including ERP systems, plant management information systems, facility management systems, process control systems, product lifecycle management systems, warehouse management information systems, energy management systems, and manufacturing execution systems. In some companies, CECM data items may even be contained in digital twins for products, production processes, and factories. Achieving an efficient extraction of relevant CECM data items from these systems through existing data exports and transformation functions is part of the next research steps. CECM data items may also be extracted from existing documents, such as material safety data sheets, product specification documents, dangerous goods specification documents, application documents and permits from environmental authorities, and special documents required when chemical substances are involved (e.g., REACH documents). The use of NLP methods to extract CECM data items from these documents seems to be a promising approach. Clearly, the (initial) company-specific data of the repository need to accurately reflect the current CECM circumstances of the firm. Hence, when chances occur, respective new items of CECM concern need to be inserted into the repository and already existing items need to be updated or deleted.

6.3 Forthcoming method evaluation and future improvements

At the current state of this ongoing research, a relevance assessment assistance system is implemented targeting a prototype solution. The prototype is implemented based on the Python programming language. Various open source NLP packages, such as SpaCy and NLTK, are used to benefit from contained general functions for text preprocessing tasks (e.g., tokenization, stemming, and lemmatization) and general functions for NER analysis and clustering. Some guiding principles for the implementation of tailored text processing algorithms based on standard components are adopted from the subject identification method proposed by (Jamil et al. 2017).

The prototype system is used to evaluate both the proposed relevance assessment method as a whole and each of the NLP-based functions. It can be expected that the insights from the evaluation will lead to revisions. The future revised method may use more advanced NLP techniques, including elements of the BERT framework that builds on machine learning techniques. Because of the multilanguage context of the application domain, a bilingual approach may contribute to a better performance of the targeted relevance assessment method. Guidance for an extension toward measuring semantic textual similarity with bilingual word-level capabilities, for example, is given by the work of (Shajalal and Aono, 2019).

Prototype-based method evaluation using a real-world dataset. The ECKD data repository of the demonstrator is populated with domain knowledge accumulated by the researchers from the scholarly literature and from a decade of intensive collaboration with researchers around the world, consulting companies, software vendors, and authorities specializing in the field of environmental compliance management. For the repository population with company-specific data, we will use a dataset provided by our industry partner, which is a globally acting mid-size production company in Germany. The company's production sites include two sites in Germany at which pressed parts, components, and automation solutions for the automotive industry and other industries are being produced. We already received a copy of the company's regulation cadaster in August 2021. This dataset contains more than one thousand regulations and approximately 400 revisions. Additionally, the cadaster stores the relevance assessments performed by the company's CECM experts and by an external consultant. Furthermore, data about scheduled and already implemented measures to enforce compliance with new regulations and revisions are also contained. The initial dataset is complemented by further company-specific data to be obtained through interviews with the company's environmental management department and an external consulting company. The respective announcement documents for the regulations and revisions will be downloaded from the corresponding official websites. The first version of the ECKD data repository will be populated with German terms and, to some degree, corresponding English terms.

Three alternatives to acquire suitable initial term collections (i.e., dictionaries) for the text analysis methods are described above. For the first prototype, the dictionaries will be obtained by performing an NER analysis with a suitable collection of German text documents selected from the document body of German environmental law and from scholarly documents. The extraction of terms from a specific classification hierarchy as targeted by the other alternatives is considered for later evaluation phases.

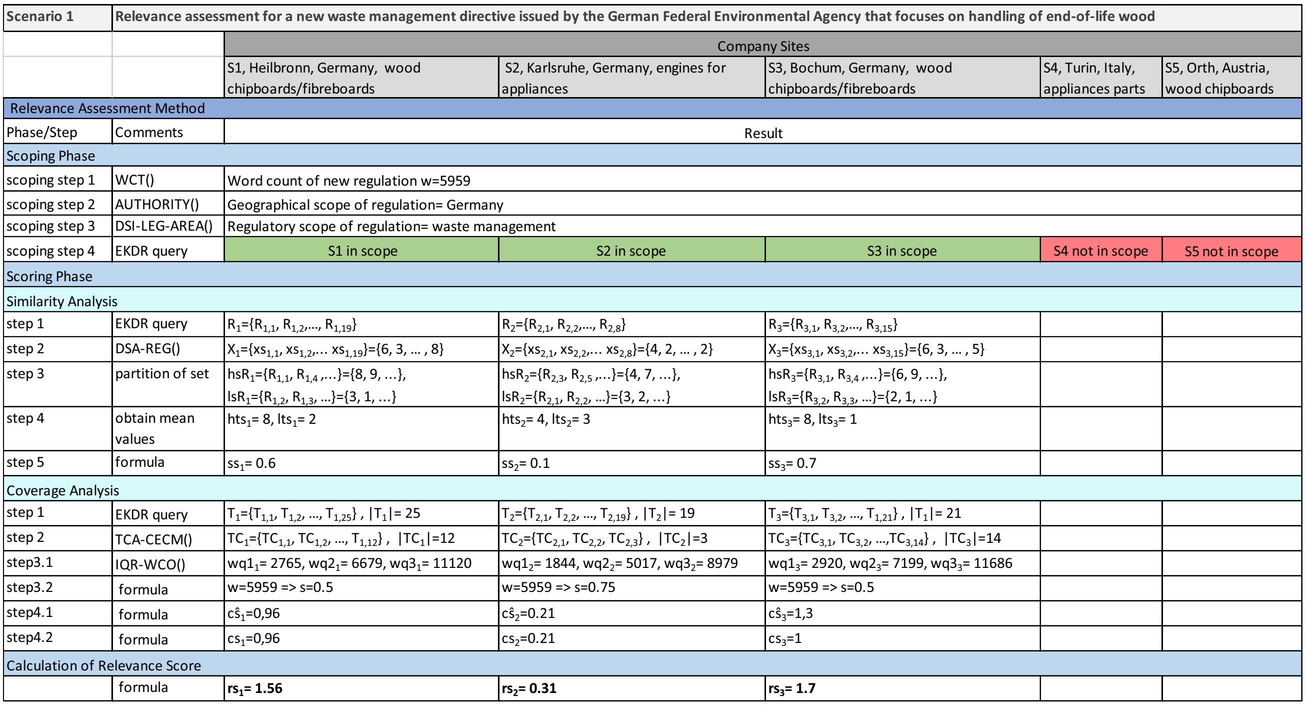
However, only part of the entire company-specific set of regulations and revisions will be populated in the initial ECKD repository. Following the general advice of machine learning practitioners, the dataset will be split into three sets: a basic system setup dataset, a validation dataset, and an evaluation dataset. The setup dataset will consist of approximately 600 assessed regulations that will be populated in the ECKD repository. Approximately 200 announcement documents will be used for a first validation of the method's accuracy. For every document, i.e., new regulation, the relevance scores for the two sites of the company will be computed with the method and compared to the relevance assessment of the CECM experts. Through this comparison, possible incorrect assessment scores of the method can be revealed. The insights obtained from the first validation can be used to improve and calibrate the method and most likely also the content of the ECKD data repository to improve the method accuracy. In a subsequent phase, the revised method will be evaluated and possibly improved again based on the evaluation dataset of approximately 200 further announcements documents assessed by the CECM experts of the company.

Considerations for future improvements of the method. The validation and evaluation of the method will most likely reveal opportunities for improvements. It is also expected that the forthcoming comparison of the method with knowledge extraction methods recently proposed for the construction engineering domain (Hassan et al. 2021; Hassan and Le, 2020, 2022; Moon et al. 2022) will yield optimization options. Several opportunities for improvements that are worth further investigation have already been identified in the present state of the research project. First, it is expected that the accuracy of the method can be improved through the use of dictionaries that are generated from existing classification hierarchies. In particular, we will investigate options to obtain dictionaries from the classification system of the environmental reporting standards, GRI (GRI 2022), and the EUROVOC topic hierarchy (Filtz et al. 2019), which is addressed, and the European law documents. A second objective of our future research is to investigate how the curation of the content could be automated through a mechanism that may involve interactions with expert users. A third improvement option concerns the text similarity analysis of the method. In the initial version of the method, the similarity analysis only considers regulations that have the same regulatory scope as the new regulation. The similarity analysis could be extended to also consider the regulations that are referred by the regulations that have the same scope. This approach to improve the method accuracy and further approaches explored in the evaluation phase will be investigated in our future research. We also plan to add a component that supplies the user with an explanation of the resulting score (i.e., a scoring report) and analytical capabilities to obtain insights about the scoring steps.

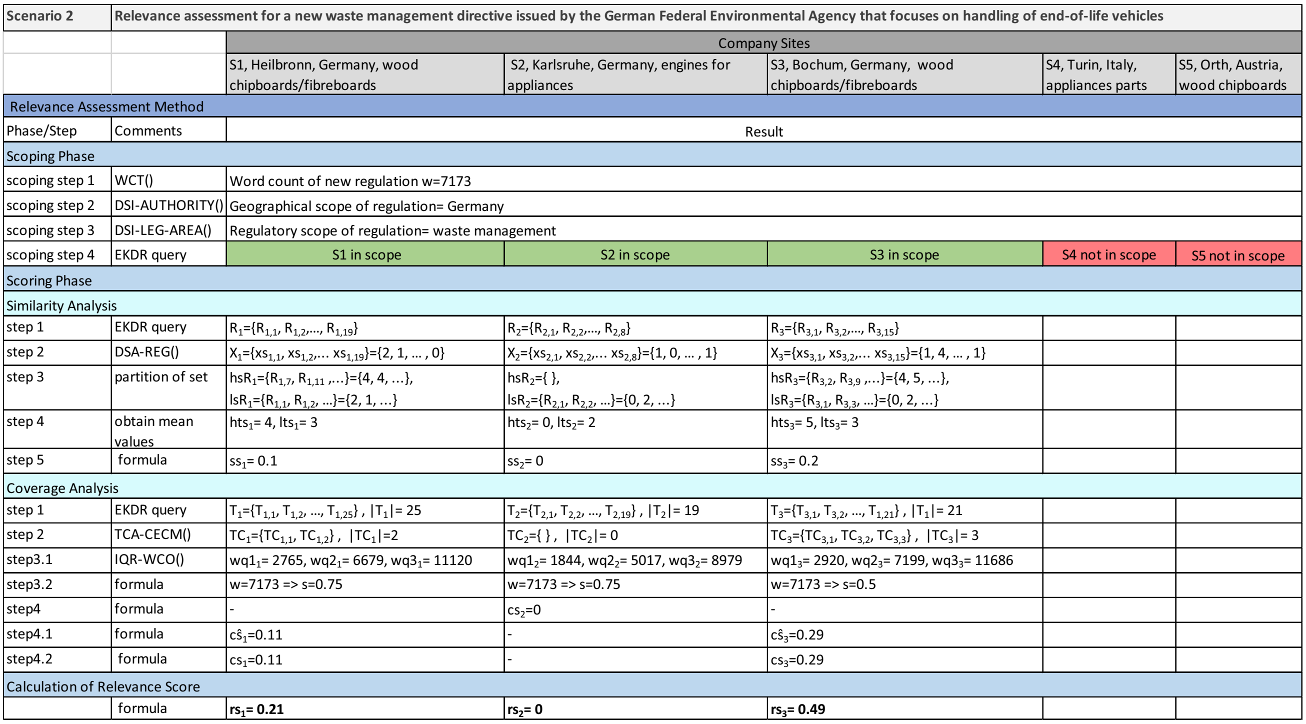
7 Conclusion

Intelligent assistance systems are already in use or are being developed for many different domains. However, it seems that today, there is still little interest in the research community and the software industry to invent and study assistance systems for corporate environmental compliance management. With the relevance assessment method and the underlying data model, core building blocks for a novel CECM assistance system are proposed in this work. It is assumed that in particular, a cloud-based approach where the domain knowledge and the company-specific data are shared by a set of assistance tools may enable companies to effectively and efficiently perform environmental compliance management duties and to prevent accidental breaches of environmental laws.

Appendix



Scenario 1	**Relevance assessment for a new waste management directive issued by the German Federal Environmental Agency that focuses on handling of end-of-life wood**					
		Company Sites				
		S1, Heilbronn, Germany, wood chipboards/fibreboards	S2, Karlsruhe, Germany, engines for appliances	S3, Bochum, Germany, wood chipboards/fibreboards	S4, Turin, Italy, appliances parts	S5, Orth, Austria, wood chipboards
Relevance Assessment Method						
Phase/Step	Comments	Result				
Scoping Phase						
scoping step 1	WCT()	Word count of new regulation w=5959				
scoping step 2	AUTHORITY()	Geographical scope of regulation= Germany				
scoping step 3	DSI-LEG-AREA()	Regulatory scope of regulation= waste management				
scoping step 4	EKDR query	S1 in scope	S2 in scope	S3 in scope	S4 not in scope	S5 not in scope
Scoring Phase						
Similarity Analysis						
step 1	EKDR query	$R_1=\{R_{1,1}, R_{1,2},\ldots, R_{1,19}\}$	$R_2=\{R_{2,1}, R_{2,2},\ldots, R_{2,8}\}$	$R_3=\{R_{3,1}, R_{3,2},\ldots, R_{3,15}\}$		
step 2	DSA-REG()	$X_1=\{xs_{1,1}, xs_{1,2},\ldots xs_{1,19}\}=\{6, 3, \ldots , 8\}$	$X_2=\{xs_{2,1}, xs_{2,2},\ldots xs_{2,8}\}=\{4, 2, \ldots , 2\}$	$X_3=\{xs_{3,1}, xs_{3,2},\ldots xs_{3,15}\}=\{6, 3, \ldots , 5\}$		
step 3	partition of set	$hsR_1=\{R_{1,1}, R_{1,4} ,\ldots\}=\{8, 9, \ldots\}$, $lsR_1=\{R_{1,2}, R_{1,3}, \ldots\}=\{3, 1, \ldots\}$	$hsR_2=\{R_{2,3}, R_{2,5} ,\ldots\}=\{4, 7, \ldots\}$, $lsR_2=\{R_{2,1}, R_{2,2}, \ldots\}=\{3, 2, \ldots\}$	$hsR_3=\{R_{3,1}, R_{3,4} ,\ldots\}=\{6, 9, \ldots\}$, $lsR_3=\{R_{3,2}, R_{3,3}, \ldots\}=\{2, 1, \ldots\}$		
step 4	obtain mean values	$hts_1= 8$, $lts_1= 2$	$hts_2= 4$, $lts_2= 3$	$hts_3= 8$, $lts_3= 1$		
step 5	formula	$ss_1= 0.6$	$ss_2= 0.1$	$ss_3= 0.7$		
Coverage Analysis						
step 1	EKDR query	$T_1=\{T_{1,1}, T_{1,2}, \ldots, T_{1,25}\}$, $\lvert T_1\rvert= 25$	$T_2=\{T_{2,1}, T_{2,2}, \ldots, T_{2,19}\}$, $\lvert T_2\rvert= 19$	$T_3=\{T_{3,1}, T_{3,2}, \ldots, T_{1,21}\}$, $\lvert T_1\rvert= 21$		
step 2	TCA-CECM()	$TC_1=\{TC_{1,1}, TC_{1,2}, \ldots, T_{1,12}\}$, $\lvert TC_1\rvert=12$	$TC_2=\{TC_{2,1}, TC_{2,2}, TC_{2,3}\}$, $\lvert TC_2\rvert=3$	$TC_3=\{TC_{3,1}, TC_{3,2}, \ldots,TC_{3,14}\}$, $\lvert TC_3\rvert=14$		
step3.1	IQR-WCO()	$wq1_1= 2765$, $wq2_1= 6679$, $wq3_1= 11120$	$wq1_2= 1844$, $wq2_2= 5017$, $wq3_2= 8979$	$wq1_3= 2920$, $wq2_3= 7199$, $wq3_3= 11686$		
step3.2	formula	w=5959 => s=0.5	w=5959 => s=0.75	w=5959 => s=0.5		
step4.1	formula	$c\hat{s}_1=0,96$	$c\hat{s}_2=0.21$	$c\hat{s}_3=1,3$		
step4.2	formula	$cs_1=0,96$	$cs_2=0.21$	$cs_3=1$		
Calculation of Relevance Score						
	formula	$\mathbf{rs_1= 1.56}$	$\mathbf{rs_2= 0.31}$	$\mathbf{rs_3= 1.7}$		



Scenario 2	Relevance assessment for a new waste management directive issued by the German Federal Environmental Agency that focuses on handling of end-of-life vehicles				
		Company Sites				
		S1, Heilbronn, Germany, wood chipboards/fibreboards	S2, Karlsruhe, Germany, engines for appliances	S3, Bochum, Germany, wood chipboards/fibreboards	S4, Turin, Italy, appliances parts	S5, Orth, Austria, wood chipboards
Relevance Assessment Method						
Phase/Step	Comments	Result				
Scoping Phase						
scoping step 1	WCT()	Word count of new regulation w=7173				
scoping step 2	DSI-AUTHORITY()	Geographical scope of regulation= Germany				
scoping step 3	DSI-LEG-AREA()	Regulatory scope of regulation= waste management				
scoping step 4	EKDR query	🟩 S1 in scope	🟩 S2 in scope	🟩 S3 in scope	🟥 S4 not in scope	🟥 S5 not in scope
Scoring Phase						
Similarity Analysis						
step 1	EKDR query	$R_1=\{R_{1,1}, R_{1,2},\dots, R_{1,19}\}$	$R_2=\{R_{2,1}, R_{2,2},\dots, R_{2,8}\}$	$R_3=\{R_{3,1}, R_{3,2},\dots, R_{3,15}\}$		
step 2	DSA-REG()	$X_1=\{xs_{1,1}, xs_{1,2},\dots xs_{1,19}\}=\{2, 1, \dots, 0\}$	$X_2=\{xs_{2,1}, xs_{2,2},\dots xs_{2,8}\}=\{1, 0, \dots, 1\}$	$X_3=\{xs_{3,1}, xs_{3,2},\dots xs_{3,15}\}=\{1, 4, \dots, 1\}$		
step 3	partition of set	$hsR_1=\{R_{1,7}, R_{1,11},\dots\}=\{4, 4, \dots\}$, $lsR_1=\{R_{1,1}, R_{1,2}, \dots\}=\{2, 1, \dots\}$	$hsR_2=\{\ \}$, $lsR_2=\{R_{2,1}, R_{2,2}, \dots\}=\{0, 2, \dots\}$	$hsR_3=\{R_{3,2}, R_{3,9},\dots\}=\{4, 5, \dots\}$, $lsR_3=\{R_{3,1}, R_{3,3}, \dots\}=\{0, 2, \dots\}$		
step 4	obtain mean values	$hts_1= 4$, $lts_1= 3$	$hts_2= 0$, $lts_2= 2$	$hts_3= 5$, $lts_3= 3$		
step 5	formula	$ss_1= 0.1$	$ss_2= 0$	$ss_3= 0.2$		
Coverage Analysis						
step 1	EKDR query	$T_1=\{T_{1,1}, T_{1,2}, \dots, T_{1,25}\}$, $\lvert T_1\rvert= 25$	$T_2=\{T_{2,1}, T_{2,2}, \dots, T_{2,19}\}$, $\lvert T_2\rvert= 19$	$T_3=\{T_{3,1}, T_{3,2}, \dots, T_{1,21}\}$, $\lvert T_1\rvert= 21$		
step 2	TCA-CECM()	$TC_1=\{TC_{1,1}, TC_{1,2}\}$, $\lvert TC_1\rvert=2$	$TC_2=\{\ \}$, $\lvert TC_2\rvert= 0$	$TC_3=\{TC_{3,1}, TC_{3,2}, TC_{3,3}\}$, $\lvert TC_3\rvert= 3$		
step3.1	IQR-WCO()	$wq1_1= 2765$, $wq2_1= 6679$, $wq3_1= 11120$	$wq1_2= 1844$, $wq2_2= 5017$, $wq3_2= 8979$	$wq1_3= 2920$, $wq2_3= 7199$, $wq3_3= 11686$		
step3.2	formula	w=7173 => s=0.75	w=7173 => s=0.75	w=7173 => s=0.5		
step4	formula	-	$cs_2=0$	-		
step4.1	formula	$c\hat{s}_1=0.11$	-	$c\hat{s}_3=0.29$		
step4.2	formula	$cs_1=0.11$	-	$cs_3=0.29$		
Calculation of Relevance Score						
	formula	**$rs_1= 0.21$**	**$rs_2= 0$**	**$rs_3= 0.49$**		

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10669-023-09900-7.

Author contributions The author confirms sole responsibility for the following: study conception and design, data collection, analysis and interpretation of results, and manuscript preparation.

Funding Open Access funding enabled and organized by Projekt DEAL. The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

Data availablity The datasets generated during and/or analyzed during the current study are not publicly available due to individual privacy concerns but are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The author has no relevant financial or non-financial interests to disclose.

References

Anandarajan M, Hill C, Nolan T (2019) Practical text analytics: maximizing the value of text data. Advances in analytics and data science:, vol 2. Springer, Cham

Bird S, Klein E, Loper E (2009) Natural language processing with Python. Sebastopol, California: O'Reilly. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=443090

Braun R, Schulz WF, Krcmar H, Russ M, Maute S, Hermann B et al (2004) System architecture and maintenance of the ecoradar web portal. In: Scharl A (ed) Advanced information and knowledge processing. environmental online communication. Springer, London, pp 147–160

Butler T (2011) Compliance with institutional imperatives on environmental sustainability: building theory on the role of Green IS. J Strateg Inf Syst 20(1):6–26. https://doi.org/10.1016/j.jsis.2010.09.006

Butler T, McGovern D (2012) A conceptual model and IS framework for the design and adoption of environmental compliance management systems. Inf Syst Front 14(2):221–235. https://doi.org/10.1007/s10796-009-9197-5

Campbell Gemmell J, Marian Scott E (2013) Environmental regulation, sustainability and risk. Sustainability 4(2):120–144. https://doi.org/10.1108/SAMPJ-Jan-2012-0003

CDP (2022) CDP Disclosure Insight Action (Home Page). https://www.cdp.net/en/

Chen P (1976) The entity-relationship model—Towards a unified view of data. ACM Trans Database Syst 1(1):9–36. https://doi.org/10.1145/320434.320440

Cyras V, Lachmayer F (2014) Compliance and Software Transparency for Legal Machines. In: Database and Information Systems: Proceedings of the 11th International Baltic Conference, Baltic

DB&IS 2014. Tallin: Tallin University of Technology Press, pp 325–336

D'hondt J, Verhaegen P-A, Vertommen J, Cattrysse D, Duflou JR (2011) Topic identification based on document coherence and spectral analysis. Inf Sci 181(18):3783–3797. https://doi.org/10.1016/j.ins.2011.04.044

Dale R (2019) Law and word order: NLP in legal tech. Nat Lang Eng 25(1):211–217. https://doi.org/10.1017/S1351324918000475

Deng Q, Hine M, Shaobo J, Sujit S (2019) Inside the black box of dictionary building for text analytics: a design science approach. J Int Technol Inf Manage 27(3):119–159

Elia F (2020) How to Compute the Similarity Between Two Text Documents? https://www.baeldung.com/cs/ml-similarities-in-text

European Union (2022) Environment and climate change: EURO-Lex home > Summaries of EU Legislation > Environment and climate change. https://eur-lex.europa.eu/summary/chapter/20.html

Filtz E, Kirrane S, Polleres A, Wohlgenannt G (2019) Exploiting EuroVoc's hierarchical structure for classifying legal documents. In: Panetto H, Debruyne C, Hepp M, Lewis D, Ardagna CA, Meersman R (eds) Lecture notes in computer science. on the move to meaningful internet systems: OTM 2019 conferences, vol 11877. Springer, Cham, pp 164–181. https://doi.org/10.1007/978-3-030-33246-4_10

Foley J, Sarwar SM, Allan J (2018) Named entity recognition with extremely limited data. http://arxiv.org/pdf/1806.04411v2

Freundlieb M, Teuteberg F (2009) Towards a Reference Model of an Environmental Management Information System for Compliance Management. In: Wohlgemuth V, Page B, Voigt K (eds), /Berichte aus der Umweltinformatik]. Environmental informatics and industrial environmental protection: concepts, methods and tools: Proceeding. 23rd International Conference Environmental Informatics (EnviroInfo), Sept. 2009, HTW Berlin, Germany. Aachen: Shaker, pp 139–148

Geissdoerfer M, Vladimirova D, Evans S (2018) Sustainable business model innovation: a review. J Clean Prod 198:401–416. https://doi.org/10.1016/j.jclepro.2018.06.240

German Environment Agency (2019) A guide to environmental administration in Germany. https://www.umweltbundesamt.de/sites/default/files/medien/376/publikationen/190722_uba_lf_environadmin_21x21_bf.pdf

Ghavami P (2020) Big data analytics methods: Analytics techniques in data mining, deep learning and natural language processing, 2nd edn. De Gruyter, Boston, Berlin

Giblin C, Müller S, Pfitzmann B (2006) From Regulatory policies to event monitoring rules: towards model-driven compliance automation. Zürich, Switzerland

GRI (2022) About GRI (Home Page). https://www.globalreporting.org/about-gri/

Gudivada VN, Rao CR (eds) (2018) Handbook of statistics / series editor C.R. Rao, C.R. Rao AIMSCS, University of Hyderabad Campus, Hyderabad, India: volume 38. Computational analysis and understanding of natural languages: principles, methods and applications. North-Holland an imprint of Elsevier, Amsterdam, Oxford

Haney BS (2019) Applied natural language processing for law practice. SSRN Electron J. https://doi.org/10.2139/ssrn.3476351

Illinois University Library (2022) Text Mining Tools and Methods. https://guides.library.illinois.edu/c.php?g=405110&p=5804542#s-lg-box-18413496

Jagota A (2020) Named Entity Recognition in NLP: Real-world use cases, models, methods: from simple to advanced. https://towardsdatascience.com/named-entity-recognition-in-nlp-be09139fa7b8

Jamil NS, Ku-Mahamud KR, Din AM, Ahmad F, Pa NC, Ishak WHW et al (2017) A subject identification method based on term frequency technique. Int J Adv Comput Res 7(30):103–110. https://doi.org/10.19101/IJACR.2017.730020

Keretna S, Lim CP, Creighton D (2014) A hybrid model for named entity recognition using unstructured medical text. In: 2014 9th International Conference on System of Systems Engineering (SOSE). IEEE. pp 85–90, https://doi.org/10.1109/SYSOSE.2014.6892468

Kerrigan SL (2003) A software infrastructure for regulatory management and compliance assistance (PhD thesis). Stanford University

Mattera P, Baggaley AK (2021) The other environmental regulators: how states unevenly enforce pollution laws. Washington, DC. https://www.goodjobsfirst.org/sites/default/files/docs/pdfs/otherregulators.pdf

Moon S, Chi S, Im SB (2022) Automated detection of contractual risk clauses from construction specifications using bidirectional encoder representations from transformers (BERT). Autom Constr 142:104465

Neto J (2021) Best NLP Algorithms to get Document Similarity. https://medium.com/analytics-vidhya/best-nlp-algorithms-to-get-document-similarity-a5559244b23b

Robinson S, Arbez G, Birta L, Tolk A, Wagner G (2015) Conceptual modeling: Definition, purpose and benefits. In: 2015 Winter Simulation Conference (WSC). IEEE. pp 2812–2826, https://doi.org/10.1109/WSC.2015.7408386

Ruhl JB (1997) Environment by making a mess of environmental law. Houst Law Rev 34(4):101–164

Schaltegger S, Freund FL, Hansen EG (2012) Business cases for sustainability: the role of business model innovation for corporate sustainability. Int J Innov Sustain Dev 6(2):95. https://doi.org/10.1504/ijisd.2012.046944

Shajalal M, Aono M (2019) Semantic textual similarity between sentences using bilingual word semantics. Prog Artif Intell 8(2):263–272. https://doi.org/10.1007/s13748-019-00180-4

Thimm H (2015) IT-supported assurance of environmental law compliance in small and medium sized enterprises. Int J Comput Inf Technol 4(2):297–305

Thimm H (2017a) ICT support of environmental compliance—Approaches and future perspectives. In: Wohlgemuth V, Fuchs-Kittowski F, Wittmann J (eds) Advances and new trends in environmental informatics: stability, continuity, innovation. Springer, Cham, pp 323–333. https://doi.org/10.1007/978-3-319-44711-7

Thimm H (2017b) Towards an intelligent assistance system to improve environmental compliance continuity. Int J Comput Inf Technol 6(5):1–8

Thimm H (2022) Systems theory-based abstractions and decision schemes for corporate environmental compliance management. Sustain Oper Comput 3:188–202. https://doi.org/10.1016/j.susoc.2022.01.007

Thimm H (2018) Towards an Active Assistance and Collaboration Support Platform for Cloud-based Corporate Environmental Compliance Management. In: Bungartz HJ, Kranzlmüller D, Weinberg V, Weismüller J, Wohlgemuth V (eds), Enviroinfo: Environmental Informatics - Techniques and Trends: Adjunct proceeding 32nd edition of the EnviroInfo: Munich, Sept. 2018, Aachen: Shaker Verlag, pp 50–55

Thimm H (2019) Investigating Website Disclosure of Corporate Environmental Compliance Management. In: Scharlach R, Simon KH, Weismüller J, Wohlgemuth V (eds) Environmental informatics: computational sustainability: ICT methods to achieve the UN Sustainable Development Goals: 33rd Conference Environmental Informatics, Adjunct Proceedings. Shaker Verlag

ul Hassan F, Le T (2020) Automated requirements identification from construction contract documents using natural language processing. J Leg Aff Dispute Resolut Eng Construct. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000379 ul Hassan F, Le T (2022) Extraction of activities information from construction contracts using natural language processing (NLP) methods to support scheduling. In: Jazizadeh F, Shealy T, Garvin MJ (eds) Construction research congress 2022. American Society of Civil Engineers, Reston, VA, pp 773–781. https://doi.org/10.1061/9780784483961.081

ul Hassan F, Le T, Lv X (2021) Addressing legal and contractual matters in construction using natural language processing: a critical review. J Constr Eng Manage. https://doi.org/10.1061/(ASCE)CO.1943-7862.0002122

WBCSD, WRI (2004) A Corporate Accounting and Reporting Standard The Greenhouse Gas Protocol: A Corporate Accounting and Reporting Standard (Revised Edition). https://ghgprotocol.org/sites/default/files/standards/ghg-protocol-revised.pdf

White R, Heckenberg D (2012) Legislation, regulatory models and approaches to compliance and enforcement: Briefing Paper No. 6