UNIVERSITY OF TWENTE.

Criteria of Information Value in Information Retrieval:

The context of Housing Corporation Risk Management

Author
Maria-Elena Tsigkou

Examination Committee
Dr. A.B.J.M. Wijnhoven
Dr. R. Klaassen

August 27th, 2020

Acknowledgements

I would like to express my gratitude to the individuals who helped in making this document possible, starting with my supervisors from the university, dr. A.B.J.M. Wijnhoven and dr. R. Klaassen, for their valuable input and support. I would also like to thank drs. M. van Grinsven, prof.dr. M.E. Iacob and Bibian Rosink for their guidance in different phases of the project. Concurrently, this project would not be possible without the support of Naris, and especially Bas van Beek and Henk Benkemper.

Finally, I would like to thank my friends and family who were there throughout the journey of my studies in Twente. Fitri, Fania, Febby, Eva, my friends and project buddies in BIT. Alex, Bram, Chris, Dilton, Dimitris, Jose, Krassi, Lida, Marilena, Nikos, Semere, Shaman, Vassilina, and Vassilis, the friends I met in Enschede along the way. The study tour gang and the D&D gang from Victor's campaign. Margarita and Raphael, my best friends from Greece who've been there from the start. Agathi, my classmate and best buddy from the moment we started our first UT course. My family; my parents, Melita and George, and my brother Peter, for their support and inspiration in life. And of course, my grandparents, Peter and Mathilde; this study would not be possible without their support and thus is devoted to them.

Abstract

Digital transformation has been a source of advancement for many a field, including Risk Management. As the information flows grow, so does the ambition to convert unstructured data into insights. The focus has been limited, however, in the pursuit of isolating information of quality. This thesis aims to investigate potential indicators of information value through the scope of housing corporation risk management. The study objectives are achieved by the examination of potential information value indicators in literature and practice. To this end, the components of a housing corporation risk tool that act as a "filter" of valuable information, and can contribute to Ontology Learning, are developed. These components include a housing corporation risk ontology, a web crawler, and a text classifier. In the implementation phase, information that is crawled from digital news sources is classified as "housing corporation risk" or "not housing corporation risk" in multiple iterations. As the model is trained, we observe successful attempts in reducing information waste, but a challenge in identifying risk-related items with high accuracy.

Table of Contents

List of Figures

List of Tables

List of Acronyms

CAS	Casualty Actuarial Society. 1, 17
COSO	Committee of Sponsoring Organizations of the Treadway Commission. 1, 3, 17
CSS	Cascading Style Sheets. 22–24, 31
CSV	Comma-separated Values. 21–23, 31
DOM	Document Object Model. 23, 30
DPA	Data Protection Authority. 26
DSRM	Design Science Research Methodology. 7
ERM	Enterprise Risk Management. 1, 17
EU	European Union. 18
GDPR	General Data Protection Regulation. 26
GRC	Governance, Risk Management, and Compliance. 2, 17
HC	Housing Corporation. 18
HCRO	Housing Corporation Risk Ontology. 19
HTML	Hypertext Markup Language. v, 22, 23, 25, 30
ICT	Information and Communication Technology. 18, 19
ISO	International Organization for Standardization. 1, 3, 18
JSON	JavaScript Object Notation. 21–23
NLP	Natural Language Processing. 13, 28
NUTS	Nomenclature of territorial units for statistics. 29
OWL	Web Ontology Language. 11, 19
RDF	Resource Description Framework. 11, 12, 19
SPYDER	Scientific Python Development Environment. 22
SVM	Support Vector Machines. 14, 28
URL	Uniform Resource Locator. 22–26, 31
WSW	Wet Sociale Werkvoorziening. 19
XML	Extensible Markup Language. 11, 21, 22
XPath	XML Path Language. 22, 23

1 Introduction

On May 27, 2017, a parking garage of 900 square meter size, located in the city of Eindhoven, and under construction at the time, collapsed due to construction error. The investigation that followed revealed that the combination of a hot day, and the uneven distribution of prefabricated concrete slabs, lead to the collapse of the fourth floor, which in a case of snowball effect lead to the collapse of the floors below. The question that arises is whether the collapse could have been avoided should there be greater risk management, information management and communication of the two.

The International Organization for Standardization (ISO) defines the concept of risk as the *effect of uncertainty on objectives* and the process of risk management as *the systematic application of management policies, procedures and practices to the activities of communicating, consulting, establishing the context, and identifying, analysing, evaluating, treating, monitoring and reviewing risk*. Concurrently, the activities of *risk identification*, *analysis* and *evaluation* are defined as risk assessment [1] [2]. Frameworks, tools and models have been gradually developed in the context of Enterprise Risk Management (ERM), such as the Committee of Sponsoring Organizations of the Treadway Commission (COSO) ERM framework or the Casualty Actuarial Society (CAS) framework. These frameworks, along with ISO 31000, 31010 and Guide 73, create an outline of an organisation's workflow concerning risk management while maintaining a general viewpoint. However, in practice, there are no universal risk classifications.

The relationship between knowledge management and risk management has not been substantial in the past since knowledge sharing can be contradictory to traditional industry standards [3]. Risk management, as a relatively young field in the academic realm, typically relies on traditional scientific methods and expert-based review in the process of risk identification [4]. However, in the era of technological transformation, new approaches are needed in the pursuit of modernising the field while fulfilling the goal of reducing uncertainty.

1.1 Motivation

The World Wide Web has seen exponential expansion in the last 30 years. By 2016, the amount of global IP traffic had been estimated by Cisco at 6.8 zettabytes, a number expected to be tripled by 2021 [5]. The "Zettabyte Era" has been facilitated by the growth of broadband speeds, mobile traffic and video streaming. A digital transformation of this size can lead to information overload for human users and missed potential for machines that cannot yet recognise unstructured data to bring insights. As the information flows grow, so does the ambition and challenge to unlock its potential.

Along the lines of this capitalisation, researchers of the scientific world, and entrepreneurs of the business world have been trying to find the optimal way to harness these vast amounts of information. Even though research is extensive in disciplines such as Information Retrieval and Information Extraction, the focus is limited regarding the criteria and metrics that should be utilised to isolate information of quality [6]. New approaches, such as text mining techniques, could act as a conduit in the process of finding and isolating meaningful data from information clutter in order to reduce uncertainty; in other words, reduce risk.

1.2 Scope

Housing corporations, also referred to as housing associations, are public or private bodies that provide affordable housing. Housing corporations in the Netherlands are private organisations, which operate under the Dutch Housing act [7]. Housing corporations own around 75% of rented dwellings in the Netherlands [8]. Since housing corporations are state-regulated, they are subject to legislation alteration as a result of economic, environmental or societal changes. Amendments of the respective legislation can have a major impact on housing corporations as well as their tenants.

Naris is a software organisation focusing on the digital transformation of Governance, Risk Management, and Compliance (GRC). Naris aims to expand their risk knowledge base by monitoring digital news sources, followed by the retrieval of news items, and the notification of clients to whom the retrieved object is relevant. Prior to the development of this service, the organisation would like to determine which criteria can be associated with the retrieval of valuable information. Awareness in regard to the factors that contribute to valuable information will allow the utilisation of the large mass of unstructured information that composes the

World Wide Web while preventing information overload.

To this end, this study investigates the possible transformation of risk management through the combination of the disciplines of information retrieval, semantic technologies and information science. Under this approach, the scope of housing corporation risk management is selected as a case study.

1.3 Problem Statement

The studies in this domain are limited in regard to a number of different levels. Firstly, we observe limited research in utilising cutting edge research in the field of Enterprise Risk Management, a field considered to be substantial in the business world and still young in the academic world [4]. Due to traditional industry standards in this field, information is private, fragmented and not standardised. Even though the COSO framework and ISO 31000 are utilised by innumerable companies worldwide, they are broad and act as mere high-level guidelines of the risk management process. Concurrently, to the best of the author's knowledge, there is no unified risk taxonomy, while the number of risk management open source tools, such as the *Open Risk Manual* [9], is limited. The second and larger facet of the problem is not domain-specific; researchers have given considerable focus on information extraction and retrieval approaches, while giving limited focus to which factors could affect the value of information [6], such as bias, time, quality and information waste.

The current situation does not facilitate the option of higher-level reasoning. Consequently, it is important to investigate potential indicators of information value, while striving to a consensus regarding the terminology of the discipline of risk. Succeeding in these objectives can expedite the transformation of the discipline in both industry and academia. Thus, the focus of this research is not on the act of retrieval itself but on the criteria that are used to assess information prior and after the retrieval. In order to operationalise the potential criteria of information value, a proof of concept of a housing corporation risk tool is developed. The components of the tool include a housing corporation risk ontology, a web crawler and a classifier that, when integrated, act as a filter of valuable information in multiple steps.

1.4 Contribution to Theory & Practice

From the perspective of academia, this research can provide insights to the blooming field of risk management, as well as the under-researched information value indicators in this context. With the creation of a housing corporation risk ontology that follows standardised terminology we encourage the promotion and expansion of open initiatives and knowledge sharing in the field of risk management.

From the perspective of practice, in connection with the collaboration with Naris, this research will directly influence Naris' development of a graph knowledge base that utilises information from external news sources, with both the use of the risk ontology, and the insights in regard to which criteria to take under consideration in order to extract information of value. The process of developing the risk ontology can act as an additional contribution, as Naris and any other stakeholder can use the approach to develop an ontology of an alternate risk domain or expand the functionality of the current ontology.

2 Research Approach

The research goal of this research has been defined as follows:

Investigate information value indicators in the context of housing corporation risk management through the use of ontology-focused crawling.

The goal can thus be divided into distinct research objectives, namely:

- The examination of information value indicators in the literature,
- the examination of information extraction approaches in the selected context,
- the examination of ontology-focused crawling approaches,
- the development of a relevant risk ontology,
- the development and execution of an ontology-focused crawler in order to evaluate the former examinations, and
- the evaluation of the risk ontology

To realise these objectives, the following research questions will be answered:

- **RQ1**: *What are the criteria that have been linked with the formulation of valuable information extracted from digital media?*
- **RQ2**: *What are the main risks of housing corporations in the Netherlands?*
- **RQ3**: *Are the criteria of RQ1 representative of reality in regard to housing corporation risks?*

2.1 Methodology

The research conducted in this report consists of a literature review and the development of an ontology, a web crawler and classifier. Throughout the execution of the tool comprised of these components, we operationalise potential criteria of information value that were found while conducting the literature review.

2.1.1 Literature Review

As a means of answering the first research question, a literature review of relevant publications was performed. The areas of interest that were considered include Information Retrieval & Extraction, Semantic Web approaches, Ontological Risk Management, Knowledge & Information Management and Philosophy of Information. These disciplines were selected to compose a rounded approach to the identification of information value indicators both in general and in the context of risk management.

Publications of the aforementioned disciplines were found through the digital academic libraries Scopus, IEEE Xplore Digital Library and Google Scholar. Focus was given to peer-reviewed information science journals such as the Journal of Information Science, the American Journal of Information Science and Technology and the Journal of the Association for Information Science and Technology. Finally, to a lesser extent, related news articles and white papers were referenced.

Among the keywords that were used in the review, some examples are "information AND retrieval AND risk", "knowledge AND graph", and "information AND value AND risk AND management". In the exploration of popular keywords, results of inapplicable disciplines were not included. For instance, results from the disciplines of medicine and biology were excluded while reviewing the keyword "knowledge base". Thus, the search terms in this instance were edited to:

TITLE-ABS-KEY (knowledge AND base) AND (EXCLUDE (SUBJAREA , "MEDI") OR EXCLUDE (SUBJAREA , "BIOC") OR EXCLUDE (SUBJAREA , "AGRI") OR EXCLUDE (SUBJAREA , "EART"))

2.1.2 Proof of Concept of Housing Corporation Risk Tool

For the objectives of developing the proof of concept, the Design Science Research Methodology (DSRM) framework is used [10], as displayed in Figure 1. This framework was selected as it provides a complete cycle of development of an artefact from a scientific perspective.

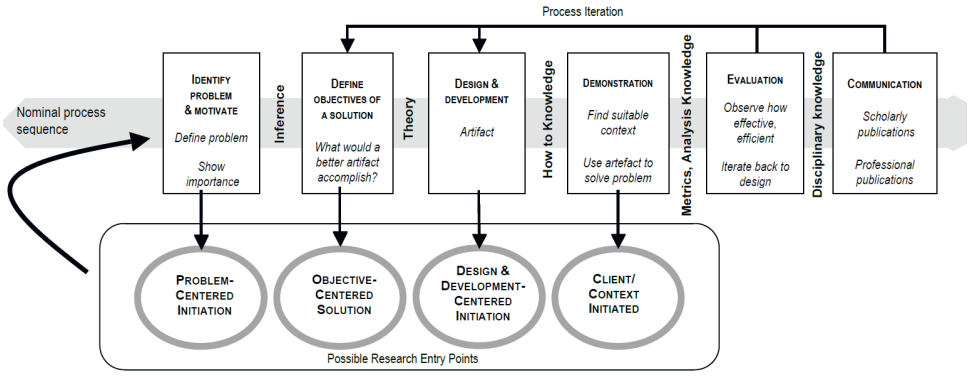

Figure 1: Design Science Research Methodology (DSRM) Process Model [10]

In summary, the contents of the domain ontology are initially used to train the classifier. The crawler is used to retrieve data from selected sources. Next, the classifier tags the relevant data, enabling us to discard the non relevant data. Finally, the relevant data are broken down to keywords that can be used to expand the ontology. An overview of this process, in conjunction with the respective stage of the DSRM framework, is displayed in Figure 2.

2.2 Structure of Report

Following the introduction and research approach in Chapters 1 and 2 respectively, Chapter 3 details a literature review of the relevant disciplines of the topic. Chapter 4 follows with a step by step description of the preparation of the development of the Housing Corporation Risk tool, while Chapter 5 presents the analysis of the results. The report concludes with Chapters 6 and 7 which present the discussion, conclusion, thoughts on potential future work, as well as the limitations of this research.

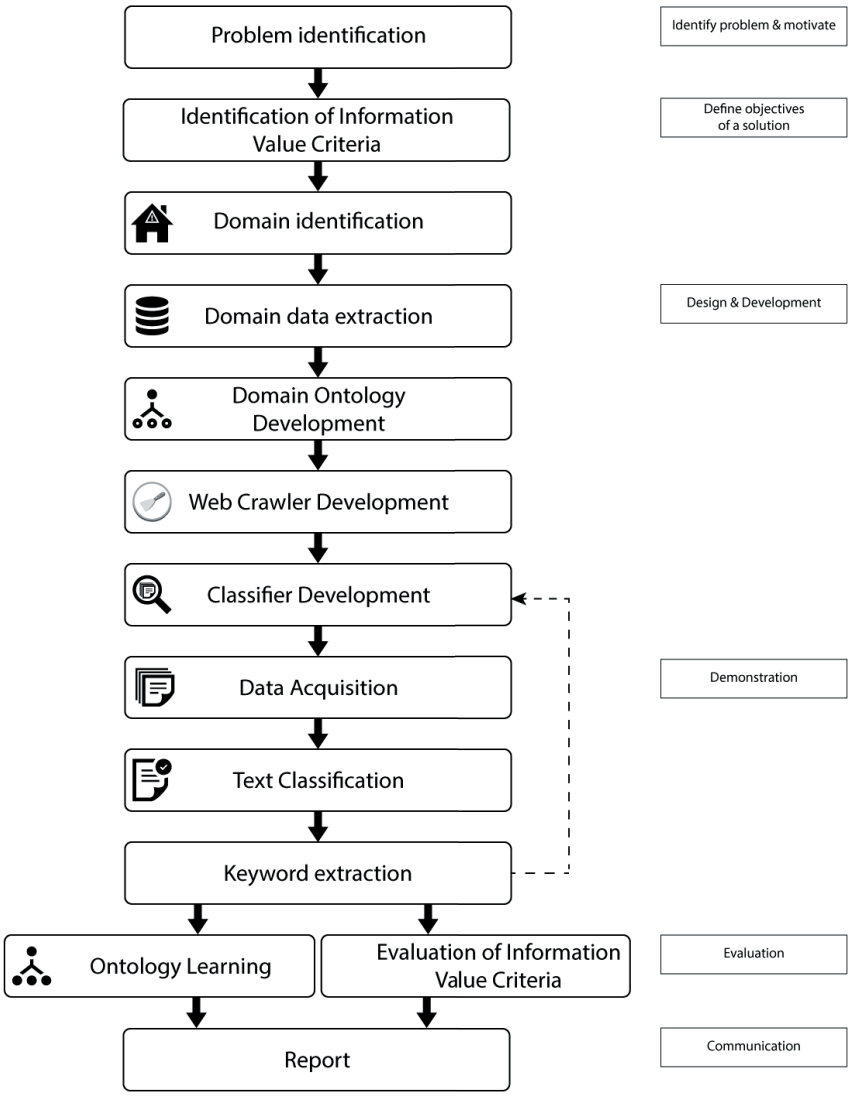

Figure 2: Approach overview

3 Criteria of Information Value in the Literature

In this chapter, the related literature will be presented. The chapter is divided into three thematic sections, referring to the background of each discipline that was investigated. The first thematic section includes a description of the terms information, knowledge and data, how they have evolved through the years, and their definition in respect to information science and its reference disciplines. The second thematic section delves into semantic information in the World Wide Web and indicates attempts of its utilisation in the field of Risk Management. The chapter concludes with an investigation of potential criteria of information value in the literature.

The pursuit of acquiring and conveying information is not new, albeit being heavily amplified with the arrival of "Big Data". Through the ages, many a philosopher has attempted to deconstruct the concept of knowledge, along with philosophical paradoxes such as "the problem of the criterion". The problem has been expressed by Chisholm [11] as two questions; *What do we know? What is the extent of our knowledge?*, and *How are we to decide whether we know? What are the criteria of knowledge?*

Instances of this reflection appear in the works of philosophers such as Michael De Montaigne and Plato who also ponder, through their individual approaches, how do we ask about an entity without knowing what it is [12] [13]. Thus, long before the entrance of "data", philosophers and researchers alike were trying to define what knowledge is and how it can be deconstructed. To date, there are attempts but no clear answer to this problem.

In the Knowledge Management, Information Science and their reference disciplines, the terms data, information and knowledge are vital. Zins lists 130 definitions of 45 scholars for these concepts [14]. Vakkari [15] notes how these terms can be taken for granted within the field of information seeking and argues that there is a paradox since their definition is, simply put, vague. Other disciplines, from communication to information science, have alternate definitions of these concepts with distinct scope or viewpoint. A consequence of this issue is the plethora of classifications that depict the types of data, information or knowledge. Each term has multiple classifications depending on the viewpoint. In statistics, for instance, we usually refer to qualitative or quantitative data. In software development, data are taxonomised as

integer, string, boolean, etc. Knowledge can be classified as tacit and explicit, with the latter being relevant to information retrieval and extraction. In 2018, following the implementation of the General Data Protection Regulation in Europe, industry and academia had to redesign their approach to data and data collection from the viewpoint of information security. In this respect, data can be categorised as public, internal or private, and restricted [16].

Buckland [17] introduced three uses of information that can be used in information science, namely Information-as-process, Information-as-knowledge and Information-as-thing. The first is described as the act of informing, the second refers to the specific "knowledge" that is communicated (a fact, subject, or event), while the third refers to what Buckland calls objects, data or documents. The author makes a connection of the latter with knowledge representation, a term that is now related to the field of Artificial Intelligence [18]. Floridi [19] argues that information can be five types of data, not mutually exclusive; *Primary*, the principal data in a database, *Secondary*, informative absent data, *Meta*, information about the Primary data, *Operational*, data about the operations of an information system, or *Derivative*, data extracted to detect patterns. In total, in a publication which introduces information as a concept, he has three distinct classifications, as depicted in Figure 3. Nevertheless, this depiction is not extensive.

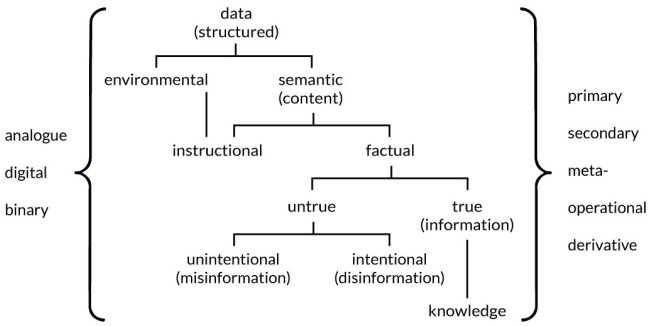

Figure 3: Floridi's map of information concepts

In information science, two common categorisations are the structured (e.g. databases) and unstructured (e.g. plain text, audio) [20]. On the web, the types of data or information exist in many formats. However, between the myriads of classifications it is hard to find a high-level extensive taxonomy of the types and formats of documents that can be found online. The margins between data, information and knowledge are not always definitive [19] [21]. In this study, we use Floridi's *General Definition of Information* to define the term, as displayed in

Table 1.

Table 1: Floridi's General Definition of Information (GDI) [19]

σ is an instance of information, understood as semantic content, if and only if:
1) σ consists of n data, for ≥ 1;
2) the data are well-formed;
3) the well-formed data are meaningful.

3.1 Information in the World Wide Web

The World Wide Web has redefined how information is approached and consumed. The conditions that facilitate the evolution and expansion of information on the web, namely its decentralised nature, along with the ease of publishing content at low costs, can be viewed as both its strengths, as well as its weaknesses.

3.1.1 The Semantic Web

Ever since the inception of the World Wide Web, its creator, Sir Tim Berners-Lee, has had an idea of what the next version would be. He described the "Semantic Web" as an extension of the Web which gives "well-defined meaning" to information [22].

Semantic networks have been defined as "graph structure[s] for representing knowledge in patterns of interconnected nodes and arcs" [23]. The first model was introduced in the 1960s by Allan M. Collins, M. Ross Quillian and Elizabeth F. Loftus, a cognitive scientist, a linguist and a psychologist, respectively. In the present, experts in the field are not optimistic about the complete success of such an endeavour, mainly due to the human variable. In a study by Anderson and Rainie [24], 895 experts were asked about the realisation of Berners-Lee's vision by the year 2020. Challenges that were identified by the experts include the fact that user-generated content is not tagged properly, that the average user will not really notice it, and that machines don't understand natural language that well yet. Some recipients referenced the human "lazy" factor, namely the fact that people do not always provide accurate descriptions and they may lie. The consensus is that even if the semantic web becomes a reality, it will not be fully implemented by 2020 and may have a different form than what Berners-Lee imagined.

Presently, the realisation of the semantic web is associated with components and standards such as the Resource Description Framework (RDF), the Extensible Markup Language (XML), the Web Ontology Language (OWL) and more [25]. The common denominator of these

components is that they are created to be readable by humans and machines alike by returning structured data.

Another form of knowledge representation is an ontology, defined in information science as an explicit specification of a conceptualisation [26]. The construction of an ontology is useful in formally representing domain knowledge, while enabling the understanding of the structure of information between both software agents and people [27]. An ontology is comprised of *classes*, *properties* and *instances*. Classes are the concepts described in the domain in question. Classes have subclasses based on the specificity of the concept. For instance, a subclass of the class "risk" is a "finance risk". The class hierarchy is structured through a top-down approach (general to specific concepts), a bottom-up approach (specific to general concepts), or a combination of the two. The internal structure of the classes is described by properties. Finally, instances are the most specific concepts of the ontology.

3.1.2 Knowledge Graphs & Linked Data

Knowledge Graphs and Linked Data are interrelated technologies that are also associated with the implementation of the Semantic Web. According to Ehrlinger and Wöß[28], knowledge graphs *acquire and integrate information into an ontology and apply a reasoner to derive new knowledge*. The authors investigated the concept of knowledge graphs since it has been widely used in academic and business environments alike, but is still unclear, and at times confused with knowledge bases and ontologies. The term "Knowledge Graph" was coined in the 1980s by researchers from the universities of Groningen and Twente but became popular, and confusing for the field, when Google presented a construct in 2012 with the same name. The definitions prior to [28] range. For instance, the definition of the Journal of Web Semantics specifies the inclusion of relationships between entities that populate the graph [29], while the definition of Farber et al. [30] explicitly mentions the Resource Description Framework (RDF) which is described as a graph-based data model used to structure and link data that describe things in the world [31]. A popular project that can be described as a knowledge graph is DBpedia, a dataset of extracted information from Wikipedia containing more than 2.6 million entities [32]. Linked data are machine-readable structured data that can be linked with similarly structured datasets [31].

3.1.3 Applications of the Semantic Web in Risk Management

Researchers have taken an interest in the utilisation of the Semantic Web to revamp risk management activities [33]. Ding et al. note that in construction risk management identical information may be presented differently since experts identify information individually. Hence, the utilisation of semantic information could be a solution. Sheth [34] singles out the sectors of finance and government and proposes a semantic approach to mitigate the complexity that follows scoring information from multiple sources. Wu et al. [35] focused on the integration of data in a knowledge graph in order to interpret the actions of Quality Assurance Directors in high-risk cases. Finally, Pittl, Fill and Honegger [36] created an ontology for risk and mitigation measures.

3.2 Information Retrieval

The state of the open web, being composed of billions of web pages that are structured disparately, complicates the operation of information retrieval techniques. Information retrieval (IR) is the discipline, and text mining technique, that focuses on the retrieval of unstructured material, usually in the form of text, from a large selection of stored data [20].

The process of an information retrieval system, such as a web crawler or a search engine, typically begins with the selection of a set of hyperlinks. The order of retrieval is set to breadth-first (retrieving each depth level sequentially), depth-first (retrieving by depth and backtracking) or by an alternate algorithm, such as PageRank. The retrieved information is stored and can be indexed and ranked using distinct criteria, such as popularity. The results can then be returned to the user in a ranked list.

Techniques that facilitate the analysis of data that were acquired through information retrieval include information extraction, Natural Language Processing (NLP), text summarization, text classification and clustering [37] [20].

Information Extraction (IE) refers to the extraction of meaningful information from a large corpus. The extracted information includes attributes that specify relationships within a corpus. NLP refers to the automatic processing of unstructured text. Clustering refers to the classification of text in groups based on the similarity of terms or patterns.

Text summarization includes text processing techniques such as tokenization, stop word re-

moval, and stemming. The process of tokenization includes the division of the retrieved text into words, referred to as *tokens*, and the removal of unnecessary characters such as punctuation and white spaces. Tokens can be normalised in order to be matched as keywords. For instance, the terms *book* and *Book* belong in the same set and are thus grouped in the same equivalence class. A similar technique is stemming, which reduces terms to their basic form. For instance, the terms *book* and *books* can be reduced to *book*. Finally, stop word removal is the elimination of common words that are not considered to be keywords in the domain in question. Text summarization techniques differ between distinct languages. For instance, stop word removal in English includes words such as *the, a, on, that* etc., while stop word removal in Dutch includes *de, en, van, ik, te* etc.

Text classification refers to the process of assigning a text object to a set of pre-determined classes. In the approach of machine learning based text classification a set of data is trained and the classification rules are learned automatically. Popular statistical models used for text classification include the Naïve Bayes, Support Vector Machines, Logistic Regression and Neural Networks. The Naïve Bayes probabilistic learning method uses Bayes's Theorem in order to predict text categories. A Support Vector Machines (SVM) is a vector space based machine learning method that explores the boundaries between two classes by representing pieces of text as points in a multidimensional space. The points that are mapped close to each other are then assigned to a category. Linear Regression is a statistical method that predicts a value based on a set of features. Finally, Deep Learning refers to an approach that emulates the way the human brain processes information through the use of artificial neural networks.

3.3 Retrieval of Valuable Information

Approaches such as the Semantic Web, can facilitate and enhance information-seeking tasks such as information retrieval and information extraction. There has been significant research in these fields in regard to the optimal way of retrieving or extracting information respectively. Little focus is given, however, to the criteria that make information of quality or of value [38] [6]. One of the main challenges that ensue, as the amount of user generated content found on the web is unparalleled, is the lack of quality control [39] [40].

When we speak of information quality, we may refer to the instance of information itself or the

quality of the source, which can in turn refer to the web page / publisher or the author. Rieh [39] labels these as *institutional level of source* and *individual level* respectively. The author focused on quality and authority and argued that users will judge the quality of information based on the authority of the respective source. Zhu [38], after testing six metrics (*currency*, *availability*, *information-to-noise ratio*, *authority*, *popularity*, and *cohesiveness*), argues that metrics of information quality can improve search effectiveness. They found that *information-to-noise ratio* can be adopted as a metric to assess the quality of information. Wijnhoven, Dietz and Amrit [41] investigated a similar concept in the context of website quality; information waste. The authors list the metrics *access speed*, *number of incoming links*, *number of broken links*, *currency* and *frequency of access* as information waste indicators. Knight and Burn [6] assembled the information quality frameworks that have been developed by researchers and found that the most common "dimensions" are *accuracy*, *consistency*, *security* and *timeliness*.

In the context of Risk Management, the identification of valuable digital information is an area where few researchers have focused. While there is research in the field of risk information quality, it is limited and fragmented between individual domains of risk management. Amir and Lev [42] remark that in the accounting domain, financial data alone cannot always provide value-relevant information. The authors investigate the cellular industry and identify industry-specific non-financial indicators such as *population*, *penetration rate* and *churn rate*. Sajko, Rabuzin and Bača [43] attempted to define information value in the context of security risk assessment. Similarly to Amir and Lev [42], they argue that financial values are not the sole influence. The authors derive to a model of three dimensions; namely *meaning to the business*, *cost defining* and *time*. Other sectors are more time-sensitive; Arsevska et al. [44] inspect disease outbreak in the health domain. They argue that while detection of relevant information is getting more complicated due to the growing amount of data, it is beneficial to use automated approaches of biosurveillance to be ahead of a possible outbreak. In the context of information security management, through an online survey, Shamala [45] identified *accuracy, amount of data, objective, completeness, reliability* and *verifiability* as information quality criteria.

A list of evaluated criteria that have been identified as potential value indicators by researchers is presented in table 2.

Table 2: Criteria of Information Value

Criteria	Publications
Accessibility	[46], [43]
Bias, Lack of	[41], [40]
Industry	[42], [44], [33], [34]
Language	[44], [47]
Quality of instance	[6], [46], [39], [43], [45], [34], [41], [38]
Quality of source	[46], [39], [47], [45], [34], [41], [38]
Quantity	[46], [43], [38]
Relevance	[44], [33], [47], [34], [41]
Space	[44], [18], [47]
Time	[44], [43], [34], [41], [38]
Waste, Lack of	[39], [43], [34], [41], [38]

4 Design & Development of Tool Components

In this chapter, the development of a domain ontology of housing corporation risks, the development of a web crawler using the Scrapy framework, and the development of a text classifier is described. The three artefacts compose aspects of the operationalisation of the information value criteria displayed in Table 2.

4.1 Housing Corporation Risk Ontology (HCRO)

Ontology-based tools of information extraction can utilise an ontology that was created manually by an expert of the domain in question to extract relevant data [48].

4.1.1 Risk Classifications and Ontologies in Business & Academia

Enterprise Risk Management (ERM) may be populated with frameworks such as the Committee of Sponsoring Organizations of the Treadway Commission (COSO) ERM framework or the Casualty Actuarial Society (CAS) framework but in practice there are no universal risk classifications. In point of fact, no consensus has been reached on the categorisation of risks within organisations [49]. Each framework has each own upper-level taxonomy, such as Hazard, Financial, Operational and Strategic, as defined by CAS. COSO references Strategic, Operational, Financial and Compliance as the "typical" risk categories. The World Economic Forum which releases a yearly report of "Global Risks" classifies risks as Economic, Environmental, Geopolitical, Societal and Technological.

In the academic dimension, the ontology-based approaches are focused on specific risk domains. Gonzalez-Conejero et al. [50] developed an ontology in relation to legal and automatic compliance of organisations in Spain within the field of Governance, Risk Management, and Compliance. In the context of disaster management, Baučić et al. [51] developed the EPISECC ontology, while Murgante et al. [52] specified even more, into the seismic domain. Hofman [53], Emmenegger et al. [54] and Palmer et al. [55] developed supply chain risk ontologies. In the housing sector, researchers have worked on ontologies regarding construction costs [56]. To the best of the author's knowledge, there is no existing ontology in the context of housing corporations' risks. However, some governments have reports with basic taxonomies of risks in this sector, usually in the form of yearly reviews. Due to the lack of universality in risk identification, each organisation, Naris included, apply their own data.

4.1.2 Methodology

The ontology was developed using the Noy and McGuinness [27] Knowledge-Engineering methodology. Initially, the domain of the ontology was selected, namely Housing Corporation risks. The process of development of the ontology is displayed in Figure 4.

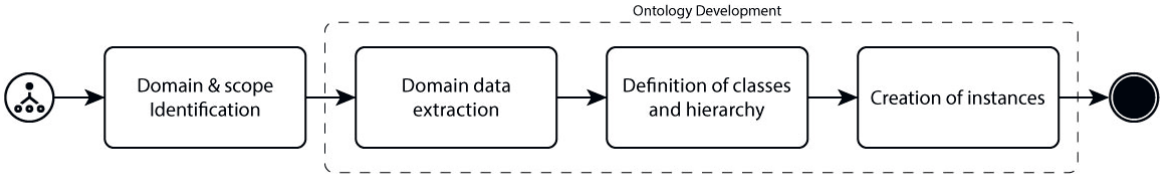

Figure 4: Process of developing the Housing Corporation Risk Ontology

Since an ontology for housing corporation risks does not exist publicly, the next step was the term enumeration by compiling a list of housing corporation risk events. To this end, data relevant to housing corporations were extracted from the database of Naris. The extraction included distinct risk events, their IDs in the database, and the risk category with which they are associated. The data were then translated from Dutch to English. The extraction included 259 risk events out of which 44 were discarded as duplicate entries. Concurrently, a glossary of risk concepts such as *Consequence*, *Risk Source* and *Likelihood*, was prepared in accordance with the Naris database and the definitions provided in ISO Guide 73:2009 [1].

Table 3: Sample of HC risks in English and Dutch

Risk Event (English)	Risk Event (Dutch)
Asbestos contamination in one or more homes	Asbestbesmetting in een of meerdere woningen
Decline in property value	Daling in waarde van het onroerend goed
Non-compliance with EU regulations by the organization	Niet tijdig voldoen aan EU-regelgeving door de organisatie
Unauthorized persons have access to ICT systems	Onbevoegden hebben toegang tot ICT systemen

A limitation of the list of HC risks is that it is not exhaustive. Compiling an exhaustive list is outside of the scope of this project since this achievement would have to be developed over a larger time frame. Simultaneously, the extracted data from the Naris database are not linked with individual client cases with respect to their privacy. The final list of 215 risk events was classified in 23 distinct categories from the Naris database. A sample of the risk events is displayed in Table 3, while the risk categories by Naris are displayed in Table 4.

The next step consisted of the definition of classes and the configuration of their hierarchy in a bottom-up approach. More specifically the bottom-level concepts in the ontology are the instances of the class *RiskEvent*, such as "Credit management is not adequate". The

Table 4: Risk Categories in English and Dutch

Risk Category in English	Risk Category in Dutch
Management & Maintenance	Beheer & Onderhoud
Finance	Financiën
Real estate development	Vastgoedontwikkeling
Human Resources & Organisation	Personeel & Organisatie
Rental	Verhuur
Activity Outsourcing / supplier management	Uitbesteding van activiteiten / leveranciersmanagement
Sales	Verkoop
Neighborhood development	Wijkontwikkeling
External communications	Externe communicatie
Collaboration	Samenwerking
Supervision	Toezicht
Purchasing & Tenders	Inkoop & Aanbesteding
Working Conditions	Arbeidsomstandigheden
Staff development	Personeelsontwikkeling
Fraud	Fraude
Facility Affairs	Facilitaire Zaken
Contract management	Contractbeheer
Management & Organization	Management & Organisatie
Complying with legislation / Compliance with internal and external regulations	Voldoen aan wetgeving / naleven interne- en externe regelgeving
Information and Communication Technology (ICT) / Automation	Informatie- en communicatietechnologie (ICT) / Automatisering
Strategy and policy development	Strategie- en beleidsontwikkeling
WSW Business risks	WSW Business risks

top-level class hierarchy of the HCRO includes *Thing*, the most general class of the ontology, which expands to four classes, namely *RiskDomain*, *RiskCategory*, *RiskEvent* and *RiskVariable*. The class *RiskDomain*, indicating the risk domain in question, contains the subclass *Housing Corporations*. The last type of entities of the ontology are the object and data properties which define associations between the classes and subclasses or provide additional information. For instance, the object property *hasCategory* defines the relationship between the classes *RiskEvent* and *RiskCategory*. The cardinality in this case is set as single, indicating that a *RiskEvent* can only have one *RiskCategory*.

The development of the ontology was performed via the tool "Protégé", originally developed by the Stanford University School of Medicine. Protégé is an open source platform that supports the latest Web Ontology Language (OWL 2) and Resource Description Framework (RDF) specifications in accordance to the World Wide Web Consortium.

4.2 Housing Corporation Risk Crawler

A web crawler is an application in the field of information retrieval, also referred to as a spider, scutter, or bot, which crawls the web and returns a collection of data. One of the typical objectives of web crawling is gathering data for search engines which will then be indexed and searched [57]. Crawlers are also used as means of digital preservation in web archiving projects, with tools such as Heritrix developed for this purpose [58]. Other types include Research Crawlers, such as CiteSeer [59], and Focused Crawlers, which target pages based on a set of topics [60]. Castillo [57] classifies crawlers as Research, Focused, Archive,

General, News Agents, and Mirroring Systems. The author taxonomises these types citing three factors, namely intrinsic quality, representational quality and freshness, arguing that, for instance, Research and Focused crawlers are more interested in the intrinsic quality whereas News Agents and Mirroring Systems are adjacent to freshness.

The process of development of a web crawler is displayed in Figure 5.

Figure 5: Process of developing the Housing Corporation Risk Crawler

4.2.1 Tool Selection

The process of identifying the optimal tool for the development of the crawler included a comparison of more than 20 web crawlers, namely: *Apache Nutch*, *Beautiful Soup*, *Bobik*, *Cheerio*, *Crawljax*, *Datahut*, *Diffbot*, *Heritrix*, *import.io*, *Mozenda*, *Octoparse*, *OutWit Hub*, *ParseHub*, *Portia*, *Promptcloud*, *Puppeteer*, *Scrape.it*, *Scraper*, *Scrapesimple*, *Scrapinghub Platform*, *Scrapy*, *UiPath*, *VisualScraper*, *Webhose.io* and *WebScraper*.

The web crawlers were evaluated by the author in accordance to the proposed functional features by Manning, Raghavan and Schütze [20], namely *robustness* and *politeness*, along with *scalability*, *efficiency*, *freshness* and *extensibility*. In detail:

- *Robustness* ensures that the crawler will be able to avoid spider traps, whether intentional or not, that may lead to a loop of requesting and fetching infinite pages. Should such a loop occur, the crawler could cause extensive load to the receiving server [61] which could result to Denial of Service, a situation where the volume of requests exceeds the response speed of the server [62]. Thus, the crawler could unintentionally disrespect the web server policies concerning the frequency of the permitted requests and disrupt the web server services in the process.

- *Politeness* refers to respecting the aforementioned web server policies by ensuring that the crawling requests are within the allowed rates of each website. In practice, these policies are specified in the Robots exclusion standard of each website, commonly known as *robots.txt*, or within HTML pages through the use of the meta tag *nofollow* [62].

- *Scalability* facilitates the customisation of the scaling crawl rate, allowing the scaling up of future work
- *Efficiency* refers to the adept use of system resources such as network bandwidth and processing power.
- *Freshness* refers to the ability to extract a new version of a previously fetched web page or document, especially in scenarios of continuous crawling.
- *Extensibility* ensures that the crawler has a modular architecture allowing moderate compatibility with new technologies such as new web protocols, formats and design methods.

In addition to the functional features, the following features were taken under consideration to select the optimal tool regarding the requirements of the study and the available resources of the researcher.

- *Software support*, referring to the operating system that the tool can be installed on. The tools that were favoured were those that support Windows or need no installation by either being cloud-based or a browser plugin in order to be supported by the resources available to the author.
- *Release history*, referring to the frequency of releases along with the date of the latest release. Tools that were favoured were those that are systematically updated, and their latest release was in 2018 and afterwards, to ensure security and compatibility with the latest release of the programming language they are based on.
- *Tool maturity*, referring to the status of stability of the tool, along with the availability of documentation and an active community.
- *Type of license* referring to tools that are open source or proprietary. Tools that were favoured were those that are open source or freeware, provided that their capabilities were within the requirements of the study.
- *Export Options*, referring to the capability of exporting data in applicable formats such as CSV, XML or JSON.

The evaluation of the criteria above resulted in the selection of two tools, Octoparse and Scrapy. Octoparse is a visual web crawler that utilises a "point and click" approach where the user clicks the elements to be scraped and the tool applies a machine learning algorithm to locate the data, meta-data and markup tags of the element. Scrapy is one of the most widely used open source crawling tools. It is Python-based, has vast documentation, and allows additional customisability and thus scalability [63]. At the time of selection (Spring 2019), both tools had recent updates, with their latest release being in November 2018 and January 2019 respectively. Even though Octoparse has a short learning curve, it provides limited functionality in comparison with Scrapy. Therefore, Scrapy was chosen as the software that will be utilised to develop the Housing Corporation Risk Crawler.

4.2.2 The Scrapy Framework

Scrapy is an application framework for website crawling, designed to extract structured data from specified unstructured sources [63]. Scrapy utilises two types of HTML selectors, CSS and XPath selectors, to extract data from multiple web pages and can generate exports to CSV, XML and JSON. The web crawler script is composed of the files *items.py* , *middlewares.py*, *pipelines.py*, *settings.py* and the directory *spider* to place the spiders of the project. The spiders are created by the user and constitute the main script. The spiders are built in the Scientific Python Development Environment (SPYDER)), an open source software that facilitates data analysis in Python. The requests are generated by a spider and the *Scrapy Engine* executes the crawl, by scraping the URLs. The response from the web server is a copy of the requested HTML elements of the web page, which are then stored as *Items* and exported in the requested format, either locally or in the cloud.

HTML Selectors

Scrapy uses XPath and CSS selectors to locate and extract HTML elements. The two types of selectors can be used jointly or independently. XPath selectors use XML Path Language (XPath), a language that selects nodes in XML documents as well as HTML documents. CSS selectors locate and extract HTML elements per their CSS stylesheet language, instead of XML nodes. Even though Scrapy supports both methods, the crawler operates with XPath expressions in any case, as CSS selectors are converted to XPath selectors. Table 5 displays a code snippet regarding the extraction of text within the CSS class of a list (li) entitled "next".

Even though the syntax of the two methods is distinct, the crawler returns identical results.

Table 5: HTML Selectors

Method	Snippet
XPath Selector	response.xpath("//li[@class='next']/text()").get()
CSS Selector	response.css('li.next a::attr(href)').get()
Combination of methods	response.css("li.next a").xpath("@href").get_all()

4.2.3 Developing the Crawler

Prior to the development of the spiders, the websites that will be crawled, also referred to as the *seed set*, have to be investigated to identify the distinct HTML and CSS elements that will be targeted to prepare the requirements and restrictions of each source. Each website that was selected for this study was inspected. The distinct HTML markup tags, classes and IDs from the Document Object Model structure of the web pages were catalogued in order to be incorporated in the crawler script.

The crawler performs four tasks, namely extracting data from specified static URLs, repeating the extraction by following the pagination of each URL (if it is applicable), following the extracted URLs of the articles and extracting the respective content within the *body* tag, and saving the extracted data into a CSV or JSON file. In this study, the CSV format is used.

Scrapy provides an interactive shell entitled *Scrapy shell* which facilitates instantaneous testing of CSS or XPath expressions for a specific URL. This feature was used to customise the query in regard to the HTML elements that have to be located in order to extract the data from each website domain. In particular, in order to extract the titles and URLs of a list of articles on the main page of a news source, the class that contains these elements has to be identified and implemented in the tailored code. For instance, the text behind the headline of an article can be defined with a range of heading meta tags (*h1* to *h6*), while being hidden behind HTML markup tags such as *div*, *span* and *a*, or specific CSS classes of said tags.

Tailoring the spiders

For each website domain, the static URLs that are crawled refer to either the main page of the website, or an alternate URL that points to a chronological list of news items. These pages, contain data that will be extracted and data that will be ignored by the crawler. For instance, the URLs of articles will be extracted, but the URLs of menu items and advertisements will

not. Concurrently, the URLs may be located in different sections of the web-page.

In the case of NOS.nl, one of the websites that are crawled in this study, the main page, *https://nos.nl*, contains seven sections. The body of the page begins with the section *Top-stories* that has a distinct *li* class, followed by the section *Featured stories* on the left side of the wrapper, and a widget with the *Latest stories* on the right side of the wrapper. Below, there is a *Most viewed videos* section and a *News in a picture* section. The body ends with a block of news that are sorted by category, followed by the footer of the page. Out of these sections, the content that needs to be extracted is included in Topstories, Featured stories, Latest stories and the categorised stories. Upon inspection, the CSS classes and IDs that will have to be included are *#featured*, *#latest-all* and *#category-news*. A sample of the inspected script is displayed in Figure 6.

Figure 6: Sample of NOS.nl script (Homepage, section Top Stories)

Depending on the structure of each page, additional inspection may be needed to avoid pagination restrictions. A limitation of Scrapy is that it cannot always extract data of news items that are loaded dynamically with JavaScript. Additionally, it is common practice to have different pagination methods between the landing page and the blog-style pages of a website.

Scrapy Items

The data and meta-data that are extracted include the title, URL and content of each news item. Additionally, the article's timestamp and list of categories are extracted, if provided by the news source. The first step to transforming the extracted data from unstructured to structured information is to utilise Scrapy's *Item* class. Item objects are containers that parallel Python dictionaries. An Item, similarly to dictionaries, maps sets of keyword arguments as objects. Items are declared in a separate file by specifying the item class name and the individual field objects that populate it. The five fields that are extracted in this study are declared within the *HousingCorpItem*, as displayed in Listing 1. In this item, the fields

title, *url* and *content* will include primary information, which will be obligatory, while the fields *timestamp* and *category* are optional fields considering that some news sources do not provide the respective data. The field *url* is extracted from the starting URL, while the fields *title*, *content*, *timestamp* and *category* are extracted from the list of article URLs that the crawler follows.

Listing 1: Scrapy Item

```
class HousingCorpItem(scrapy.Item):
    # Primary Fields
    title = scrapy.Field()
    url = scrapy.Field()
    content = scrapy.Field()
    # Optional Fields
    timestamp = scrapy.Field()
    category = scrapy.Field()
```

Text Normalization

Having defined the HTML elements that will be extracted, the next step in the preparation of the script is ensuring that the data to be extracted will be normalised upon retrieval. To this end, the text must be isolated from HTML tags and attributes. In the example of NOS.nl, extracting the article titles with the command *response.css(".list-items ::text")* would return the following:

```
<Selector xpath="descendant-or-self::*[@class and contains
(concat(' ', normalize-space(@class), ' '), ' list-items ')]/
descendant-or-self::text()" data='Cruiseschip botst in
Venetie op kade en '>
```

Thus, Scrapy and Python commands such as *.extract()* and *.strip()* are used to remove the HTML selector information, along with characters such as \n and redundant space characters from the string. The output is then stripped to *'Cruiseschip botst in Venetië op kade en toeristenboot'*. Finally, the file *settings.py*, is edited to ensure that the crawler obeys the robots.txt rules, to configure the maximum concurrent requests permitted and to provide a description of the User-Agent.

Following Hyperlinks

There are three types of hyperlinks that the spider will have to extract and follow, namely:

- **Article URLs.** A list of URLs that point to pages of individual articles, necessary to extract the main content of each news item.

- **A pagination URL.** A single URL that, if applicable, points to the next page of listed articles. In Listing 2, if a "next-page" is detected, the function *callback=self.parse* is deployed and thus the spider repeats the initial parsing process for that page and extracts additional news items. This process can be restricted with the use of *'DEPTH LIMIT'* as a Scrapy custom setting. This parameter limits the depth level of the crawled pages by the spider. The depth limit can be tailored to each website in accordance to crawling criteria such as frequency, amount of articles in a single page, quantity of desired data and time.

- **A redirection URL.** Following the guidelines of GDPR, every digital news source in Europe that tracks cookies needs to have user consent to do so. Thus, upon first visit websites are requesting acceptance of cookies through various methods such as pop-up plugins. In some cases, the crawler can be blocked by a "cookie wall", restricting access to the content of the web-page unless consent is given. Even though cookie walls were deemed by the Dutch Data Protection Authority as non-compliant with the regulation due to the lack of freedom of choice[64], some Dutch websites continue to use them. Depending on the design of the cookie wall, some can be bypassed in the form of following a redirection URL.

Listing 2: Pagination Loop

```
next_page = response.xpath('//*[@id="next-page"]/a/@href')
.extract_first()
        if next_page:
            yield scrapy.Request(response.urljoin(next_page),
            callback=self.parse)
```

4.3 Housing Corporation Risk Classifier

The third component that was developed prior to the implementation of the crawler, was a text classifier for risks of housing corporations. The process of development of the text classifier is displayed in Figure 7.

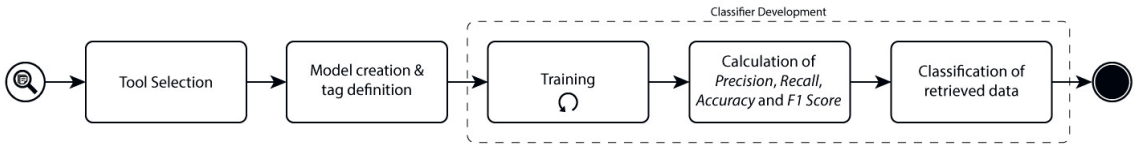

Figure 7: Process of developing the Housing Corporation Risk Classifier

4.3.1 Tool Selection

The model was created in "MonkeyLearn", an AI platform with natural language processing capabilities that facilitates the creation of multi-language custom text classifiers. The platform has a tiered pricing system, starting with a free level with restrictions on the number of custom models and queries. There is no limitation, however, on the data imported to train the model. The platform was selected due to its compatibility with Scrapy, as well as the data science application Rapidminer which is used at a later stage of the implementation of the HCR tool.

4.3.2 Training the Model

A custom classifier in the platform can be built through topic classification, sentiment analysis (e.g. positive, neutral or negative) or intent classification. Topic classification, which was selected for this model, classifies based on a topic, aspect or relevance and can be used to organise items in accordance with their subject. The text data will be classified based on the custom tags that will be defined. In the case of the HCR model, the classification criterion was whether a piece of information is a housing corporation risk or not. In other words, the classification defines if a text item is relevant to housing corporation risks. Thus the tags *housing-corporation-risk* and *not-housing-corporation-risk* were defined. A second multi-class model was built using tags based on subcategories of housing corporation risks, such as *maintenance*, *fraud* and *working-conditions*. Two datasets were then used to initially train the model; internal data, extracted from the instances of the HCR Ontology, and external data, extracted from random risk related articles, both set in the English language. The data can be imported through file transfer, directly from third party apps or through the respective

add-on in the Scrapy cloud.

In total, 330 text items were used to train the English model with a supervised approach before the first implementation. Manning [20] argues that in a scenario of limited initial data a high bias classifier is preferred. Thus, the classifier was initially trained with the use of the Multinomial Naïve Bayes. In the following iterations, Support Vector Machines were used. Concurrently, the platform can be instructed to utilise Natural Language Processing and techniques such as stemming, normalising weights, and filtering stop words. Due to the limited initial data and the complexity of the tags, the confidence levels of the second classifier were not as high as the ones of the first. In particular, confidence levels of the second ranged from 5% to 10%, while the first reached 95% or more. As a result, the second classifier was not used during the implementation phase. Following the creation of the English classifier, this process was repeated in order to create the Dutch classifier.

For each tag of the classifier, MonkeyLearn generates a keyword List and a keyword cloud, along with data on the true positive, true negative, false positive and false negative text items which compose the *precision* and *recall* of the tag and the *accuracy* and *F1 score* of the corpus. Precision refers to the number of instances that were correctly classified to a tag divided by all the correct classifications. Recall is the number of instances that were correctly classified to a tag divided by the total number of instances in that tag. The former represents how correct the classification is, while the latter represents how complete it is. The F1 score refers to the harmonic mean of the precision and recall. Finally, accuracy is the percentage of instances that were correctly predicted in their respective tag.

5 Implementation

In this chapter, a pilot implementation of the HCR tool is analysed. The implementation is described in three steps, throughout which, we examine the criteria of Table 2 to prepare the data acquisition, and use metrics to evaluate the training of the classifiers.

5.1 Preparation

The structure of the crawler was built using the websites *quotes.toscrape.com*, provided by Scrapy as a development example, and *nos.nl*, a Dutch news source. Upon completion of the first working spider for *nos.nl*, the preparation of the implementation phase was initiated with the identification of the content to be extracted. The code was then replicated and tailored to detect the characteristics of each web-page.

5.1.1 Identification of Scope

The scope of this study is information on risk for housing corporations in the Netherlands. This statement restricts the criteria of *Language*, *Space* and *Industry*. Regarding *Language*, the information to be retrieved is available in Dutch and English. Thus, we select digital news sources with the Dutch country code top-level domain (*.nl*). Concerning *Space*, in this implementation, we select news sources at a national level. In a different instance, where the scope is at a sub-national level, the selected news sources could include region-specific domains, such as *Tubantia.nl* which has a news stream targeted to the geographical area of Twente. The European classification scheme *Nomenclature of territorial units for statistics (NUTS)*[65] can be utilised to classify news sources in respect to the criterion of Space. An example is presented at Table 6. Finally, even though *Industry* is defined as housing corporation risk management, we will not target industry-specific news sources to avoid the possibility of *Bias*.

Table 6: Examples of Dutch news sources classified per location level

Location Level	News Source
National	NOS.nl
NUTS Level 1 (Lands)	NHnieuws.nl (Regional news from North Holland)
NUTS Level 2 (Provinces)	RTVoost.nl (Regional news from Overijssel)
NUTS Level 3 (COROP regions)	Tubantia.nl (Regional news from Twente)

5.1.2 Digital News Sources Selection

In a 2019 report published by the Reuters Institute and Oxford University, the Netherlands placed 4th between 38 markets regarding trust in the news [66]. According to the study, the most trusted news sources for Dutch users are *NOS News*, *RTL News*, *NU*, *AD (Algemeen Dagblad)* and *De Volkskrant*. These news sources (in Dutch), along with two news sources that publish articles in English (*dutchnews.nl* and *nltimes.nl*) were selected as the allowed domains of the crawler. Two criteria of value that are implicated in the selection of news sources are the Quality and Bias. In this case, both terms can refer to the quality/bias of the source or a specific instance. However, the complexity of identifying the quality or bias of a specific instance is high and not always feasible due to resource and time constraints. Thus, we evaluate the two criteria based on the credibility of the source, and in this case, based on user review. Concurrently, we include more than one source of information.

5.1.3 DOM Structure Inspection

The user interface, along with the HTML DOM model of each page, were investigated to extract the corresponding paths that had to be targeted. One news source, namely AD.nl, was excluded from the source list since the crawler could not bypass its cookie wall. This restriction, directly affects the criterion of *Accessibility*, as any references to this domain are directly eliminated. In total, six domains were crawled from ten spiders built to access three types of web-pages, namely main pages, category pages, and search pages. The spider list is displayed in Table 7.

Table 7: Spiders and their respective target types of pages

Spider	Target Page	Additional Settings
DutchNewsNLmain	Main , Category	
DutchNewsNLalt	Search	Pagination
NLTimesMain	Main, Category	Pagination
NOSmain	Main	
NOScat	Category	
NOSsearch	Search	Pagination
NUmain	Main, Category	
RTL	Main, Category	Pagination
deVolkskrant	Category	Redirection
deVolkskrantSearch	Search	Pagination, Redirection

Each domain required one or more distinct spider formations due to their diversity in design. In every domain other than RTL, at least one of the three types of pages that were crawled,

were designed differently. *Dutchnews.nl* has pagination in the search page, but not the main. Concurrently, there is no hyperlink provided with a complete news feed. Thus, in a project of a larger scale, the frequency of crawling would be affected as the domain would have to be scraped every few hours or days to track its content accurately. *Nltimes.nl* includes pagination in the main page. However, there multiple sections of overlapping news feeds that have to be filtered to avoid duplicate entries. The search page of this domain was the sole instance where the crawler was forbidden to access, with the following message being displayed in the error log:

```
2019-05-20 22:42:58 [scrapy.downloadermiddlewares.robotstxt]
DEBUG: Forbidden by robots.txt:
<GET https://nltimes.nl/search/node/housing?page=1>
```

NOS.nl does not have a pagination link in the main or any category page. In the search page, there is a JavaScript button that provides more results. In this case, the crawler was able to detect the pagination links. An issue unique to this domain was that a recent time stamp is extracted from the *time* attribute as, for instance, "Vandaag, 18:18" (translating to "Today, 18:18"), deeming the piece of meta-data not useful for future analysis. *NU.nl* has no pagination in the main page and an Ajax automatic loader in the category pages. The search page utilises JavaScript that rendered the article URLs unreachable for the crawler. Finally, *de Volkskrant* was the sole domain that used the same CSS classes in all types of pages. Its main and category pages utilise an automatic loader, whereas the search page has traditional pagination. Out of the six domains, none were entirely consistent in the design of the distinct types of pages.

5.2 Data Acquisition

The device where the prototype of the crawler was executed was a Lenovo laptop, with processing power of 2.6GHz and physical memory of 16GB. The depth limit of each spider was set to 15, allowing them to go through 15 pages of pagination. An additional spider was developed utilising Scrapy's *CrawlerProcess* method to allow the spiders to run concurrently. Finally, the parameter *'FEED-URI'* was added to the custom settings of each spider, instructing them to save the scraped data in a CSV file. Running the six spiders (one spider per domain) concurrently, resulted in extracting 860 news items, including a total of 224544 words from

the respective articles. The websites were crawled in less than 30 seconds. No thematic restrictions were placed in the crawler such as keywords, or pages of specific categories.

Data Acquisition was divided in two parts, due to the bilingual nature of the study. Thus, two sets of data were created, a set in English and a set in Dutch. For instance, upon execution of the crawler in regard to the English corpus, 570 distinct items were initially extracted, with 322 originating from *dutchnews.nl* and 248 from *nltimes.nl* respectively.

5.3 Text Classification

The first evaluation of the HCR English and Dutch classifiers occurred with a testing set of 100 random articles out of the extracted data. The data were then used to improve the iterations that followed.

5.3.1 First Iteration

English Classifier: Out of the 99 articles, 71 were classified correctly. The classifier was successful in predicting what was not a risk for housing corporations with 100% success rate, but failed to predict what was, with only 12.5% of these predictions being true positive. In this instance, the classifier returned 32 articles as possible indicators of house corporation risk, while it should have returned 4. The performance of the classifier is summarised in Table 8.

Table 8: Text classification in English: First execution

Classification	Texts	Precision	Recall	TP	TN	FP	FN	Accuracy	Harmean
hcr	32	12,5%	100,0%	4	67	28	0	71,7%	22,2%
not hcr	67	100,0%	70,5%	67	4	0	28	71,7%	82,7%

Dutch Classifier: In the case of the Dutch classifier, the accuracy of the results was independently evaluated by the author and an employee from Naris. Out of the 100 articles, 75 were classified correctly. Similarly to the English classifier, it was successful in predicting what was not a risk for housing corporations with 97.3% success rate, but failed to predict what was, with 8% precision. The performance of the classifier is summarised in Table 9.

In the first iteration, the two classifiers had analogous results. The classifiers have very high performance in determining which articles are not relevant to housing corporation risks. However, in detecting relevant articles there were multiple false positives, resulting in high

Table 9: Text classification in Dutch: First execution

Classification	Texts	Precision	Recall	TP	TN	FP	FN	Accuracy	Harmean
hcr	4	8,0%	50,0%	2	73	23	2	75,0%	13,8%
not hcr	96	97,3%	76,0%	73	2	2	23	75,0%	85,4%

percentages of information waste. In the end, out of the 100, only 4 and 2 articles respectively were relevant to housing corporations. This low number was expected since the crawl did not have any keyword specifications in regard to risk. In the case of the Dutch classifier, the relevance of 3 news items was considered "debatable" and had to be confirmed by expert review within Naris.

5.3.2 Second Iteration

Prior to the second iteration, the data of the first retrieval process were used as relevance feedback to improve the training of the model.

English Classifier: Out of the 100 articles, 90 were classified correctly. The classifier tagged 76 out of 82 articles successfully as not risks for housing corporations, reaching a precision of 95%. Concurrently, the was a notable improvement in predicting housing corporation risks, with a precision of 70% compared to the 12.5% of the previous iteration. The performance of the classifier is summarised in Table 10.

Table 10: Text classification in English: Second execution

Classification	Texts	Precision	Recall	TP	TN	FP	FN	Accuracy	Harmean
hcr	18	70,0%	77,8%	14	76	6	4	90,0%	73,7%
not hcr	82	95,0%	92,7%	76	14	4	6	90,0%	93,8%

Dutch Classifier: In contrast to the English classifier, and due to the lack of volume of housing corporation risk related articles, the results of the Dutch classifier were not greatly improved. In this iteration, out of 100 news items, 9 were deemed irrelevant since the crawler extracted information from hyperlinks that were advertisements. From the remaining 91 items, 23 were predicted to be relevant to housing corporation risks and 68 were predicted to be non relevant. The actual true positives were 3 and true negatives were 66. The performance of the classifier is summarised in Table 11.

Table 11: Text classification in Dutch: Second execution

Classification	Texts	Precision	Recall	TP	TN	FP	FN	Accuracy	Harmean
hcr	5	13,0%	60,0%	3	66	20	2	75,8%	21,4%
not hcr	95	97,1%	76,7%	66	3	2	20	75,8%	85,7%

5.3.3 Iterations Three to Five

English Classifier: Three more executions followed for the English classifier, adding 465 articles to the data in total. Out of the three, the highest precision in the prediction of housing corporation risks occurred in iteration 3. It also had the highest amount of confirmed housing corporation risks out of all the iterations, totalling to 16. The performance of the classifier is summarised in Table 12.

Table 12: Text classification in English: Third, fourth and fifth execution

#	Classif.	Texts	Precision	Recall	TP	TN	FP	FN	Accuracy	Harmean
3	hcr	16	78,9%	93,8%	15	80	4	1	95,0%	85,7%
	not hcr	84	98,8%	95,2%	80	15	1	4	95,0%	97,0%
4	hcr	3	60,0%	100,0%	3	60	2	0	96,9%	75,0%
	not hcr	62	100,0%	96,8%	60	3	0	2	96,9%	98,4%
5	hcr	13	58,8%	76,9%	10	180	7	3	95,0%	66,7%
	not hcr	187	98,4%	96,3%	180	10	3	7	95,0%	97,3%

5.3.4 Trained Model & Keyword Extraction

Having completed the former iterations for both classifiers, we are able to calculate the F1 score, as well as the accuracy of the trained models. Regarding the English classifier, both the overall accuracy and the F1 score are equal to 93%. For the Dutch classifier, the respective metrics are at 78%.

Table 13: Top keyword output of the English and Dutch trained models

Keywords	
English	**Dutch**
housing market, social housing, housing rent, dutch housing, rent, rent rise, incorrect, estate, rent controlled, mortgage, recovery, quality, ict, corporations boss, unauthorized, ict housing, Amsterdam housing, housing agency, housing, market recovery	organisatie, verhuur, woningen, beheer, project, financieel, vastgoedbeheer, financieel beheer, onvoldoende, gevolg, verkopen, nieuwbouw, personeel organisatie, ict, huisvesting, ict huisvesting, uitvoering, toegang, tijdig, bestaat

6 Discussion

The first objective of this study, in the form of the first research question, was to identify which criteria can act as indicators of information value. The criteria summarised in Table 2 were identified and were operationalised through the development and implementation of a crawling tool. The criteria of Language, Industry and Space were quantified in accordance with the scope of the project. These criteria were defined from the start as two distinct languages, the domain of housing corporations, and Dutch news sources on a national level, respectively. The criteria of Quality of Source and Bias are evaluated from the point of view of user trustworthiness utilising recent data from a major survey.

In the next step, the development and output of the web crawler quantifies the criterion of Accessibility. Once the code was customised to the selected news sources, the retrieval of information had a success rate of 85%, since information from every news source other than AD.nl was retrieved in a few seconds for every crawling attempt. Information Currency (Time) and Quantity were quantified through the custom settings of the crawler in accordance to the requirements and limitations of the project. The size of the retrieved corpus was then affected by Relevance, operationalised through text classification. Having filtered the initial corpus, the level of information waste is quantified as the number of false positive news items. Last but not least, Quality of Instance can be evaluated from two perspectives. Firstly, in close relation to the Quality of Source and in accordance to the respective quality metrics as examined by Knight & Burn [6]. In this case we focus on the quality of text itself, examining factors such as spelling and grammar accuracy, objectivity, and lack of spam. Secondly, in relation to how accurate and relevant the predictions of the trained model are. This metric is very low as it is negatively correlated with the high level of information waste that was observed.

The process of developing and running the Housing Corporation Risk Ontology, Housing Corporation Risk Crawler, and Housing Corporation Risk Classifiers, was used to resolve the final two research questions. Namely, the identification of risks of housing corporations in the Netherlands and the evaluation of the criteria that composed the first research question. An output of the study, along with the filtration of news items based on the selected criteria, is a keyword list (Table 13 & Figure 8) that can be used for ontology learning. Even though the corpus of each classifier was relatively small, the results indicate that the HCR tool can dispose of irrelevant articles with very high accuracy. The results of relevant articles were not

as satisfactory, even though a significant improvement was noticed in the English classifier. Time, in combination with the specificity of the domain, resulted in a limited amount of true positive results for the tag "housing corporation risks".

Figure 8: Keyword Cloud of Dutch Classifier

The 11 initial criteria of information value were all assessed throughout the course of the study, in different levels of influence, and through different methods. In summary, these criteria were defined by either the domain of the study, user trustworthiness, expert review, or the customisation of the web crawler. The diversity between the methods constitutes one of the main challenges of quantifying the importance of each criterion. However, we can argue that in this scenario the criteria of industry, language, space, accessibility, quality and relevance have been interchangeable in the pursuit of finding valuable content. Less focus was given on bias and time. The former, since it can be interconnected with the quality of the source, and the latter since the context of the study has to do with recent news items. The criterion of time can also be affected by the frequency of the implementation of the web crawler.

The use of Scrapy was effective in the fast retrieval of information from the selected news sources. The crawler is able to extract clean data that are ready to be analysed. However, there is a lot of manual work involved in order to extract the distinct elements from different domains or even different web pages of the same domain. This fact is the major limitation of this method, as described in Chapter 7.

7 Conclusion

The retrieval of valuable information facilitates the digital transformation of a field such as risk management and provides the opportunity to enhance the decision making process of Naris' clients. The operationalisation of information value is challenging, especially with scalability in mind. Every scenario is different and in every context there are distinct metrics that are ideal to measure the criteria of value. In addition, varying interpretations and abstraction in industry standards enhance the complexity of future implementations.

This thesis provides a step by step method to operationalise the retrieval of valuable information. In the context of housing corporation risk management, we successfully managed to filter out more than 80% of non-relevant data. The model is successful in detecting which articles are not related to housing corporation risks but is challenged in reducing the number of false positives in relevant articles. The limited amount of risk related news items in each crawl of the general news sources is not adequate to train the model robustly. However, by increasing the frequency of crawling iterations over time, the classifier will gain more relevant results, and in return, organically increase the value of the tool.

7.1 Limitations

A limitation of the tool that was developed in the context of this study is that, at the current stage, the filtering of information based on the criteria of information value is not automated. Concurrently, the resources of the study limited the text classification to 600 news items.

From a technical perspective, the majority of limitations and challenges surfaced while developing the web crawler. Even though Scrapy is a robust framework for focused crawling, there is a high level of customisation for each domain. Often, different parameters had to be set for the main page, a category page and the search page of a domain. This results in an added complexity of the tool and the risk of additional customisation when the structure of the web-page is altered. A trend that was observed in the Dutch news sources that were crawled, was the lack of pagination in the news feeds. Even though this fact can be registered as a limitation, in a future implementation of this tool it can be mitigated by altering the frequency of crawling.

Throughout the development of the crawler, many instances occurred where the desired re-

sponse was returned when tested in the Scrapy Shell but was not returned as the object of a Scrapy Item. An instance of this issue was observed while developing the snippet for NOS.nl. As displayed in Table 14, in the process of extracting multiple paragraphs from the body of the page of an article, the outputs of the Shell and the Spider differed. In particular, the output of the Shell was complete, returning all the paragraphs that composed the article, while the output of the Spider was partial, returning only the first paragraph.

Table 14: Comparison of retrieval between the Scrapy Shell and Spider

Method	Input	Output
Scrapy Shell	response.css ('p.text_1TQrL1WP::text') .extract()	['De terugreis van veel bezoekers van het concert van Muse in Nijmegen verliep gisteravond onplezierig. Er reden tot diep in de nacht geen treinen tussen Arnhem en Utrecht vanwege een aanrijding. Veel concertgangers strandden daardoor in Nijmegen.', 'Op het station van Nijmegen ontstond een chaos. Veel bezoekers van het concert klaagden over de drukte. Het was voor veel mensen niet duidelijk hoe ze thuis moesten komen, zo blijkt uit berichten op Twitter.', ' sprak vanochtend met concertganger Pim Olde Hampsink uit Almelo. Ook hij had na afloop de grootste moeite om thuis te komen. Om 06.15 uur was hij nog stééds niet thuis.', ... ']
NOS Spider	yield{ 'body': response ('p.text_1TQrL1WP::text') }	2019-06-28 15:55:28 [scrapy.core.scraper] DEBUG: Scraped from <200 https://nos.nl/artikel/2291030-duizenden-fans-vast-op-station-nijmegen-na-concert-muse.html> {'body': 'De terugreis van veel bezoekers van het concert van Muse in Nijmegen verliep gisteravond onplezierig. Er reden tot diep in de nacht geen treinen tussen Arnhem en Utrecht vanwege een aanrijding. Veel concertgangers strandden daardoor in Nijmegen.'}

Additionally, there are instances where browsing a page from the web browser has a different output compared to crawling. Such a limitation can occur in the form of redirection due to a cookie wall or the form of automatically generated elements through technologies such as JavaScript. Finally, there were minor discrepancies between the use of CSS and XPath selectors.

In regard to the development of the two classifiers, the size of the corpus and time restrictions of the project did not allow further improvement, especially in the case of the Dutch model. As mentioned in Chapter 6, the amount of articles referring to housing corporation risks were limited and in result the output of the respective tag was not satisfactory.

7.2 Future Work

Even though the tool is not automated, the filtering of information through the selected criteria was successful. Thus, a possible next step is the automation of the functionalities of the tool. Concurrently, continuous training of the classifiers with the utilisation of a larger corpus is crucial to improve their precision on items that have to be classified as housing corporation risks. An improvement in this regard will reduce information waste even further.

Future applications of the tool can be scaled to dynamic platforms such as social media tools. Implications to have in mind, however, are the reliance to third party software that may be modified at any time, and the quality of user generated content.

References

[1] International Organization for Standardization. (2016). ISO Guide 73:2009 - Risk management - Vocabulary, [Online]. Available: https://www.iso.org/standard/44651.html (visited on 02/20/2019).

[2] International Organization for Standardization. (2018). ISO 31000:2018 Risk management - Guidelines, [Online]. Available: https://www.iso.org/obp/ui#iso:std:iso:31000:ed-2:v1:en (visited on 12/23/2018).

[3] E. Rodriguez and J. Edwards, "People, Technology, Processes and Risk Knowledge Sharing", *Electronic Journal of Knowledge Management*, vol. 8, no. 1, pp. 139–150, 2010. [Online]. Available: http://www.ejkm.com.

[4] T. Aven, "Risk assessment and risk management: Review of recent advances on their foundation", *European Journal of Operational Research*, vol. 253, no. 1, pp. 1–13, 2016. [Online]. Available: https://doi.org/10.1016/j.ejor.2015.12.023.

[5] "Cisco Global Cloud Index: Forecast and Methodology, 2016–2021", Cisco, 2017.

[6] S. A. Knight and J. Burn, "Developing a Framework for Assessing Information Quality on the World Wide Web", *Informing Science*, vol. 8, pp. 160–172, 2005. [Online]. Available: https://doi.org/10.28945/493.

[7] H. Koolma, "The rise and fall of credibility – A way to understand the case of the Dutch public housing sector", in *NIG Working Conference*, Rotterdam, the Netherlands, 2011.

[8] Government of the Netherlands. (2019). Housing associations, [Online]. Available: https://www.government.nl/topics/housing/housing-associations (visited on 05/20/2019).

[9] Open Risk. (2019). Open Risk Manual, [Online]. Available: https://www.openriskmanual.org/wiki/Main_Page (visited on 07/25/2019).

[10] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A Design Science Research Methodology for Information Systems Research", *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2007. [Online]. Available: https://doi.org/10.2753/MIS0742-1222240302.

[11] R. M. Chisholm, *The Problem of the Criterion*. Marquette University Press, 1973.

[12] M. Montaigne, *An Apology for Raymond Sebond*, trans. by C. Cotton, ser. Essays of Montaigne, vol. 4. New York: Gnome Press, 1910.

[13] C. Woods, *Meno*, 2012. [Online]. Available: http://dx.doi.org/10.2139/ssrn.1910945.

[14] C. Zins, "Conceptual approaches for defining data, information, and knowledge", *Journal of the American Society for Information Science and Technology*, vol. 58, no. 4, pp. 479–493, 2005. [Online]. Available: https://doi.org/10.1002/asi.20508.

[15] P. Vakkari, "Information seeking in context: A challenging metatheory", in *Proceedings of an International Conference on Information Seeking in Context*, ser. ISIC '96, Tampere, Finland: Taylor Graham Publishing, 1997, pp. 451–464. [Online]. Available: http://dl.acm.org/citation.cfm?id=267190.267221.

[16] European Union, *REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*, 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[17] M. Buckland, "Information as thing", *Journal of the Association for Information Science and Technology*, vol. 42, no. 5, pp. 351–360, 1991. [Online]. Available: https://doi.org/10.1002/(SICI)1097-4571(199106)42:5%3C351::AID-ASI5%3E3.0.CO;2-3.

[18] M. Malhotra and T. G. Nair, "Evolution of Knowledge Representation and Retrieval Techniques", *International Journal of Intelligent Systems Technologies and Applications*, vol. 07, pp. 18–28, 2015. [Online]. Available: http://dx.doi.org/10.5815/ijisa.2015.07.03.

[19] L. Floridi, *Information - A Very Short Introduction*. Oxford University Press, 2010.

[20] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[21] J. Rowley, "The wisdom hierarchy: Representations of the dikw hierarchy", *Journal of Information Science*, vol. 33, no. 2, pp. 163–180, 2006.

[22] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web", *Scientific American*, 2001.

[23] J. F. Sowa, "Semantic Networks", *Encyclopedia of Artificial Intelligence*, 1992. [Online]. Available: https://doi.org/10.1002/0470018860.s00065.

[24] J. Anderson and L. Rainie, "The Fate of the Semantic Web", Pew Research Center, 2010.

[25] World Wide Web Consortium, *Standards and drafts*, 2019. [Online]. Available: https://www.w3.org/TR.

[26] T. R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing", *International Journal of Human-Computer Studies*, vol. 43, no. 5, pp. 907–928, 1995. [Online]. Available: https://doi.org/10.1006/ijhc.1995.1081.

[27] N. F. Noy and D. L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology", Tech. Rep., 2001.

[28] L. Ehrlinger and W. Wöß, "Towards a Definition of Knowledge Graphs", in *12th International Conference on Semantic Systems*, ser. SEMANTiCS 2016, Leipzig, Germany, 2016. [Online]. Available: http://ceur-ws.org/Vol-1695/.

[29] M. Kroetzsch and G. Weikum, *Special issue on knowledge graphs*, 2015. [Online]. Available: http://www.websemanticsjournal.org/index.php/ps/announcement/view/19 (visited on 01/28/2019).

[30] M. Färber, F. Bartscherer, C. Menne, and A. Rettinger, "Linked Data Quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO", *Semantic Web Journal*, vol. 9, no. 1, pp. 77–129, 2017. [Online]. Available: http://dx.doi.org/10.3233/SW-170275.

[31] C. Bizer, T. Heath, and T. Berners-Lee, *Linked Data - The Story So Far*, 2009.

[32] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, "DBpedia - A crystallization point for the Web of Data", *Journal of Web Semantics*, vol. 7, no. 3, pp. 154–165, 2009. [Online]. Available: https://doi.org/10.1016/j.websem.2009.07.002.

[33] L. Ding, B. Zhong, S. Wu, and H. Luo, "Construction risk knowledge management in BIM using ontology and semantic web technology", *Safety Science*, vol. 87, pp. 202–213, 2016. [Online]. Available: https://doi.org/10.1016/j.ssci.2016.04.008.

[34] A. Sheth, "Enterprise applications of semantic web: The sweet spot of risk and compliance", in *Industrial Applications of Semantic Web*, Springer US, 2005, pp. 47–62.

[35] J. Wu, F. Lecue, C. Guéret, J. Hayes, S. van de Moosdijk, G. Gallagher, P. McCanney, and E. Eichelberger, "Personalizing Actions in Context for Risk Management Using Semantic Web Technologies", in *The Semantic Web – ISWC 2017: 16th International Semantic Web Conference*, 2017, pp. 367–383. [Online]. Available: https://doi.org/10.1007/978-3-319-68204-4_32.

[36] B. Pittl, H.-G. Fill, and G. Honegger, "Enabling Risk-Aware Enterprise Modeling using Semantic Annotations and Visual Rules", in *Proceedings of the 25th European Conference on Information Systems (ECIS)*, Guimarães, Portugal, 2017. [Online]. Available: https://aisel.aisnet.org/ecis2017_rp/22.

[37] R. Talib, M. K. Hanif, S. Ayesha, and F. Fatima, "Text Mining: Techniques, Applications and Issues", *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 11, pp. 414–418, 2016. [Online]. Available: http://doi.org/10.14569/IJACSA.2016.071153.

[38] X. Zhu and S. Gauch, "Incorporating Quality Metrics in Centralized/Distributed Information Retrieval on the World Wide Web", in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '00,

Athens, Greece: ACM, 2000, pp. 288–295. [Online]. Available: http://doi.org/10.1145/345508.345602.

[39] S. Y. Rieh, "Judgment of information quality and cognitive authority in the web", *Journal of the American Society for Information Science and Technology*, vol. 53, no. 2, pp. 145–161, 2002. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.10017.

[40] F. Wijnhoven, "The hegelian inquiring system and a critical triangulation tool for the internet information slave: A design science study", *Journal of the American Society for Information Science and Technology*, vol. 63, no. 6, pp. 1168–1182, 2012. [Online]. Available: https://doi.org/10.1002/asi.22622.

[41] F. Wijnhoven, P. Dietz, and C. Amrit, "Information waste, the environment and human action: Concepts and research", in *ICT Critical Infrastructures and Society*, ser. HCC: IFIP International Conference on Human Choice and Computers, Amsterdam, The Netherlands: Springer, 2012, pp. 134–142. [Online]. Available: https://doi.org/10.1007/978-3-642-33332-3.

[42] E. Amir and B. Lev, "Value-relevance of nonfinancial information: The wireless communications industry", *Journal of Accounting and Economics*, vol. 22, pp. 3–30, 1996. [Online]. Available: https://doi.org/10.1016/S0165-4101(96)00430-2.

[43] M. Sajko, K. Rabuzin, and M. Bača, "How to calculate information value for effective security risk assessment", *Journal of Information and Organizational Sciences*, vol. 30, no. 2, 2006. [Online]. Available: https://doi.org/10.31341/jios.

[44] E. Arsevska, M. Roche, P. Hendrikx, D. Chavernac, S. Falala, R. Lancelot, and B. Dufour, "Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web", *Computers and Electronics in Agriculture*, vol. 123, pp. 104–115, 2016. [Online]. Available: https://doi.org/10.1016/j.compag.2016.02.010.

[45] P. Shamala, R. Ahmad, A. Zolait, and M. Sedek, "Integrating information quality dimensions into information security risk management (ISRM)", *Journal of Information Security and Applications*, vol. 36, pp. 1–10, 2017. [Online]. Available: https://doi.org/10.1016/j.jisa.2017.07.004.

[46] F. Naumann and C. Rolker, "Assessment Methods for Information Quality Criteria", in *Conference on Information Quality*, 2000, pp. 148–162. [Online]. Available: http://doi.org/10.18452/9207.

[47] H. Saggion, A. Funk, D. Maynard, and K. Bontcheva, "Ontology-Based Information Extraction for Business Intelligence", in *The Semantic Web*, Berlin, Heidelberg: Springer, 2007,

pp. 843–856. [Online]. Available: https://doi.org/10.1007/978-3-540-76298-0_61.

[48] A. H. F. Laender, B. A. Ribeiro-Neto, A. S. da Silva, and J. S. Teixeira, "A Brief Survey of Web Data Extraction Tools", *SIGMOD Rec.*, vol. 31, no. 2, pp. 84–93, 2002. [Online]. Available: http://doi.acm.org/10.1145/565117.565137.

[49] M. W. Elliot, *Risk Management Principles and Practices*. The Institute, 2012.

[50] J. Gonzalez-Conejero, A. Meroño-Peñuela, and D. Fernández-Gámez, "Ontologies for Governance, Risk Management and Policy Compliance", in *JURIX 2011*, Vienna, Austria, 2011.

[51] M. Baucic, S. Knezic, and G. Neubauer, "The EPISECC Ontology model: Spatio-temporal ontology for disaster management", M. Gahegan, Ed., 2017.

[52] B. Murgante, G. Scardaccione, and G. Las Casas, "Building ontologies for disaster management: Seismic risk domain", in *Urban and Regional Data Management 2009*, A. Krek, M. Rumor, E. M. Fendel, and S. Zlatanova, Eds., CRC Press, 2009, pp. 259–269.

[53] W. Hofman, "Supply Chain Risk Analysis with Linked Open Data", *Frontiers in Artificial Intelligence and Applications*, vol. 229, pp. 77–87, 2011. [Online]. Available: http://doi.org/10.3233/978-1-60750-785-7-77.

[54] S. Emmenegger, E. Laurenzini, and B. Thönssen, "Improving Supply-Chain Management based on Semantically Enriched Risk Descriptions", in *Proceedings of the International Conference on Knowledge Management and Information Sharing*, ser. KMIS 2012, Barcelona, Spain, 2012, pp. 70–80. [Online]. Available: http://doi.org/10.5220/0004139800700080.

[55] C. Palmer, E. N. Urwin, A. Niknejad, D. Petrovic, K. Popplewell, and R. I. M. Young, "An ontology supported risk assessment approach for the intelligent configuration of supply networks", *Journal of Intelligent Manufacturing*, vol. 29, no. 5, pp. 1005–1030, 2018. [Online]. Available: https://doi.org/10.1007/s10845-016-1252-8.

[56] F. Abanda, B. Kamsu-Foguem, and J. Tah, "Bim – new rules of measurement ontology for construction cost estimation", *Engineering Science and Technology, an International Journal*, vol. 20, no. 2, pp. 443–459, 2017. [Online]. Available: https://doi.org/10.1016/j.jestch.2017.01.007.

[57] C. Castillo, "Effective Web Crawling", in *ACM SIGIR Forum*, vol. 39, New York, NY, USA: Association for Computing Machinery, 2005, pp. 55–56. [Online]. Available: https://doi.org/10.1145/1067268.1067287.

[58] G. Mohr, M. Kimpton, M. Stack, and I. Ranitovic, "An Introduction to Heritrix - An open source archival quality web crawler", in *Proceedings of the 4th International Web Archiving Workshop IWAW'04*, Bath, UK, 2004. [Online]. Available: http://crawler.archive.org/Mohr-et-al-2004.pdf.

[59] The College of Information Sciences and Technology, PennState. (2019). About Citeseer^x, [Online]. Available: https://csxstatic.ist.psu.edu/home (visited on 06/20/2020).

[60] S. Chakrabarti, M. van den Berg, and B. Dom, "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery", *Computer Networks*, vol. 31, no. 11, pp. 1623–1640, 1999. [Online]. Available: https://doi.org/10.1016/S1389-1286(99)00052-3.

[61] M. Koster, "Robots in the web: Threat or treat?", *ConneXions*, vol. 9, no. 4, 1995.

[62] M. Thelwall and D. Stuart, "Web Crawling Ethics Revisited: Cost, Privacy, and Denial of Service", *Journal of the American Society for Information Science and Technology*, vol. 57, no. 13, pp. 1771–1779, 2006. [Online]. Available: https://doi.org/10.1002/asi.v57:13.

[63] Scrapy. (2019). Scrapy at a glance, [Online]. Available: http://doc.scrapy.org/en/latest/intro/overview.html (visited on 08/20/2019).

[64] AP. (2019). Websites moeten toegankelijk blijven bij weigeren tracking cookies, [Online]. Available: https://autoriteitpersoonsgegevens.nl/nl/nieuws/websites-moeten-toegankelijk-blijven-bij-weigeren-tracking-cookies (visited on 06/28/2019).

[65] European Commission. (2016). Regulation (EC) No 1059/2003 of the European Parliament and of the Council of 26 May 2003 on the establishment of a common classification of territorial units for statistics (NUTS), [Online]. Available: http://data.europa.eu/eli/reg/2003/1059/2018-01-18 (visited on 06/25/2019).

[66] N. Newman, R. Fletcher, A. Kalogeropoulos, and R. K. Nielsen, "Reuters institute digital news report 2019", Reuters Institute for the Study of Journalism, 2019. [Online]. Available: http://www.digitalnewsreport.org/survey/2019/netherlands-2019/.

Appendix A: Sample of Scrapy Spider Script

```
import scrapy
from ..items import HousingCorpItem
from scrapy.crawler import CrawlerProcess

class NOScat(scrapy.Spider):
    name = 'noscat'
    start_urls = [
        'https://nos.nl/nieuws/',
    ]

    custom_settings = {
        'DEPTH_LIMIT': 15,
        'FEED_URI': "hcrc1nl.csv",
        'FEED_FORMAT': 'CSV'
    }

    def parse(self, response):
        # parsing all items
        for href in response.xpath('//*[@id="latest_in_category"]/div/div/ul/li/a/@href'):
            url = response.urljoin(href.extract())
            yield scrapy.Request(url, callback=self.parse_article_page)

        # following the next page (not applicable in this hyperlink)
        next_page = response.xpath('//*[@id="next-page"]/a/@href').extract_first()
        if next_page:
            yield scrapy.Request(response.urljoin(next_page), callback=self.parse)

    # extracting content from the article page
    def parse_article_page(self, response):
        article = HousingCorpItem()
        # Primary parameters
        article['title'] = response.xpath('//h1/text()').extract_first().strip()
        article['url'] = response.url
        article['content'] = response.xpath('//div[(@class="article_block")]//p//text()').extract()
        # Optional parameters
        article['timestamp'] = response.xpath('//time/@datetime').extract()
        article['category'] = response.xpath('//div[@class="categories_3flhrYg3"]/a/text()').extract()
        yield article
```