



UNIL | Université de Lausanne

Unicentre

CH-1015 Lausanne

<http://serval.unil.ch>

Year : 2023

Essays on External Data Sourcing

Krasikov Pavel

Krasikov Pavel, 2023, Essays on External Data Sourcing

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive <http://serval.unil.ch>

Document URN : urn:nbn:ch:serval-BIB_3B7E58D3B3453

Droits d'auteur

L'Université de Lausanne attire expressément l'attention des utilisateurs sur le fait que tous les documents publiés dans l'Archive SERVAL sont protégés par le droit d'auteur, conformément à la loi fédérale sur le droit d'auteur et les droits voisins (LDA). A ce titre, il est indispensable d'obtenir le consentement préalable de l'auteur et/ou de l'éditeur avant toute utilisation d'une oeuvre ou d'une partie d'une oeuvre ne relevant pas d'une utilisation à des fins personnelles au sens de la LDA (art. 19, al. 1 lettre a). A défaut, tout contrevenant s'expose aux sanctions prévues par cette loi. Nous déclinons toute responsabilité en la matière.

Copyright

The University of Lausanne expressly draws the attention of users to the fact that all documents published in the SERVAL Archive are protected by copyright in accordance with federal law on copyright and similar rights (LDA). Accordingly it is indispensable to obtain prior consent from the author and/or publisher before any use of a work or part of a work for purposes other than personal use within the meaning of LDA (art. 19, para. 1 letter a). Failure to do so will expose offenders to the sanctions laid down by this law. We accept no liability in this respect.



UNIL | Université de Lausanne

FACULTÉ DES HAUTES ÉTUDES COMMERCIALES
DÉPARTEMENT DES SYSTÈMES D'INFORMATION

ESSAYS ON EXTERNAL DATA SOURCING

THÈSE DE DOCTORAT

présentée à la

Faculté des Hautes Études Commerciales
de l'Université de Lausanne

pour l'obtention du grade de
Doctorat ès Sciences en systèmes d'information

par

Pavel KRASIKOV

Directrice de thèse
Prof. Christine Legner

Jury

Prof. Rafael Lalive, président
Prof. Stéphanie Missonier, experte interne
Prof. Tobias Mettler, expert externe
Prof. Sirkka Jarvenpaa, expert externe

LAUSANNE
2023



UNIL | Université de Lausanne

FACULTÉ DES HAUTES ÉTUDES COMMERCIALES
DÉPARTEMENT DES SYSTÈMES D'INFORMATION

ESSAYS ON EXTERNAL DATA SOURCING

THÈSE DE DOCTORAT

présentée à la

Faculté des Hautes Études Commerciales
de l'Université de Lausanne

pour l'obtention du grade de
Doctorat ès Sciences en systèmes d'information

par

Pavel KRASIKOV

Directrice de thèse
Prof. Christine Legner

Jury

Prof. Rafael Lalive, président
Prof. Stéphanie Missonier, experte interne
Prof. Tobias Mettler, expert externe
Prof. Sirkka Jarvenpaa, expert externe

LAUSANNE
2023

IMPRIMATUR

Sans se prononcer sur les opinions de l'auteur, la Faculté des Hautes Etudes Commerciales de l'Université de Lausanne autorise l'impression de la thèse de Monsieur Pavel KRASIKOV, titulaire d'un bachelor en management de l'Université de Lausanne et d'un master en systèmes d'information de l'Université de Lausanne, en vue de l'obtention du grade de docteur en systèmes d'information.

La thèse est intitulée :

ESSAYS ON EXTERNAL DATA SOURCING

Lausanne, le 26 juin 2023

La Doyenne



Marianne SCHMID MAST



Members of the thesis committee

Prof. Christine LEGNER

Professor, Faculty of Business and Economics (HEC),
University of Lausanne, Switzerland.

Thesis supervisor

Prof. Stéphanie MISSONIER

Professor, Faculty of Business and Economics (HEC),
University of Lausanne, Switzerland.

Internal member of the thesis committee

Prof. Tobias METTLER

Professor, Swiss Graduate School of Public Administration (IDHEAP),
University of Lausanne, Switzerland.

External member of the thesis committee

Prof. Sirkka JARVENPAA

Professor, The McCombs School of Business,
University of Texas at Austin, United States of America.

External member of the thesis committee

Prof. Rafael LALIVE

Professor, Faculty of Business and Economics (HEC),
University of Lausanne, Switzerland.

President of the thesis committee

University of Lausanne
Faculty of Business and Economics

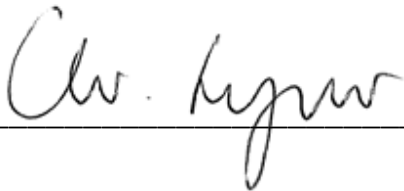
PhD in Information Systems

I hereby certify that I have examined the doctoral thesis of

Pavel KRASIKOV

and have found it to meet the requirements for a doctoral thesis.

All revisions that I or committee members
made during the doctoral colloquium
have been addressed to my entire satisfaction.

Signature:  Date: June 1, 2023

Prof. Christine LEGNER
Thesis supervisor

University of Lausanne
Faculty of Business and Economics

PhD in Information Systems

I hereby certify that I have examined the doctoral thesis of

Pavel KRASIKOV

and have found it to meet the requirements for a doctoral thesis.

All revisions that I or committee members
made during the doctoral colloquium
have been addressed to my entire satisfaction.

Signature:  Date: 1.06.2023

Prof. Stéphanie MISSIONIER
Internal member of the doctoral committee

University of Lausanne
Faculty of Business and Economics

PhD in Information Systems

I hereby certify that I have examined the doctoral thesis of

Pavel KRASIKOV

and have found it to meet the requirements for a doctoral thesis.

All revisions that I or committee members
made during the doctoral colloquium
have been addressed to my entire satisfaction.

Signature:



Date: 30.05.2023

Prof. Tobias METTLER
External member of the doctoral committee

University of Lausanne
Faculty of Business and Economics

PhD in Information Systems

I hereby certify that I have examined the doctoral thesis of

Pavel KRASIKOV

and have found it to meet the requirements for a doctoral thesis.

All revisions that I or committee members
made during the doctoral colloquium
have been addressed to my entire satisfaction.

Signature:  Date: June 1, 2023

Prof. Sirkka JARVENPAA
External member of the doctoral committee

Acknowledgments

I would like to express my sincere gratitude to the individuals and organizations who have played a crucial role in the completion of this PhD thesis. First and foremost, I extend my heartfelt appreciation to my supervisor, Christine Legner, for her exceptional guidance and expertise. Second, I would like to acknowledge the invaluable contributions of my colleagues, whose intellectual engagement and discussions have greatly influenced the development of my ideas. Third, I am indebted to my family for their support, encouragement, and understanding. Their unwavering belief in my abilities has been a constant source of inspiration throughout this journey. Finally, I would like to express my gratitude to the broader academic community for their scholarly works and research, which have shaped my thinking and contributed to the advancement of knowledge in my field.

Doctoral Thesis

ESSAYS ON EXTERNAL DATA SOURCING

Pavel Krasikov

Department of Information Systems,
Faculty of Business and Economics (HEC), University of Lausanne

Lausanne, 2023

*"The important thing is not to stop questioning.
Curiosity has its own reason for existence."*

– Albert Einstein

Abstract

In the age of digital transformation, enterprises are becoming increasingly aware of the value of external data, which originates beyond their four walls. Despite the growing number of datasets and their potential value, external data is sourced in an ad-hoc manner without clear guidelines. This leads to inconsistent sourcing decisions, characterized by a lack of clarity on the object of sourcing and the underlying data sourcing practices. Existing studies showcase scenarios of enterprises using external data, which are fraught with obstacles. A crucial challenge confronting companies that intend to use external data is to identify suitable datasets supporting specific business scenarios and to prepare them for use. In the context of a specific external data type – open data in our case – researchers have developed several data assessment techniques. Unfortunately, these techniques are limited in scope, do not consider the use context, and are not embedded in the complete set of activities required for open data consumption in enterprises. The emerging field of data sourcing also displays a notable absence of comprehensive research, prompting a clarion call for action in Information Systems (IS) research to address this gap. Considering the abovementioned research opportunities, this thesis – through three interrelated research streams – provides foundations for, analyzes, and improves data sourcing practices in the enterprise context. The first stream lays the foundations for the topic and investigates the company-wide sourcing and managing of external data. The second stream reflects on sourcing practices concerning open data, as one of the most prominent external data types, and challenges the widespread perception that open data is easily accessible and readily available. Focusing on one of the most pressing topics facing present-day companies, the third stream provides a foundation for the academic conceptualization of data sourcing in the context of sustainability.

The outcomes of this thesis project enable the transition from ad-hoc acquisition to well-informed, professional data sourcing approaches in the enterprise context. The contributions of the first research stream are an external data sourcing taxonomy (Essay 1), which informs sourcing decisions in an enterprise context, and a reference process to source and manage external data (Essay 2), which is accompanied by explicit prescriptions in the form of design principles. The second research stream proposes a use case-driven assessment of open corporate registers (Essay 3) and, building on the subsequent findings, a method to screen, assess, and prepare open data for use in support of companies' open data activities (Essay 4). Finally, the third research stream reveals and elaborates on three data sourcing practices developed by companies in response to institutional pressures in the sustainability context (Essay 5).

Contents of the thesis

Introductory paper

Essays on External Data Sourcing.....1

Essay 1

Responding to the Siren Song of External Data: A Taxonomical Approach to Data Sourcing.....43

Essay 2

Unleashing the Potential of External Data: A DSR-based Approach to Data Sourcing.....79

Essay 3

Open Data in the Enterprise Context: Assessing Open Corporate Data's Readiness for Use.....103

Essay 4

A Method to Screen, Assess, and Prepare Open Data for Use.....128

Essay 5

Introducing a Data Perspective to Sustainability: How Companies Develop Data Sourcing Practices for Sustainability Initiatives.....163

Introductory Paper on

ESSAYS ON EXTERNAL DATA SOURCING

Pavel Krasikov

Faculty of Business and Economics (HEC), University of Lausanne, Switzerland

Table of contents

- 1 Introduction..... 4
- 2 Background 6
 - 2.1 External data as an under-researched topic..... 6
 - 2.2 Emergence of a data sourcing paradigm 9
 - 2.3 Research opportunity 12
- 3 Thesis overview 13
 - 3.1 Research objectives 13
 - 3.2 Research setting 14
- 4 Research stream 1: External data sourcing 19
 - 4.1 Motivation and background 19
 - 4.2 Research objectives, methodology, and contributions 19
 - 4.3 Discussion, limitations, and outlook 21
- 5 Research stream 2: Preparing open data for use in the business context..... 24
 - 5.1 Motivation and background 24
 - 5.2 Research objectives, methodology, and contributions 25
 - 5.3 Discussion, limitations, and outlook 26
- 6 Research stream 3: Data sourcing for sustainability 28
 - 6.1 Motivation and background 28
 - 6.2 Research objectives, methodology, and contributions 29
 - 6.3 Discussion, limitations, and outlook 31
- 7 Discussion 33
 - 7.1 Implications 33
 - 7.2 Limitations..... 35
 - 7.3 Outlook and future research..... 36
- 8 References 38

List of figures

Figure 1. Overview of research streams and questions	13
Figure 2. Consortium research overview (Österle & Otto, 2010).....	17

List of tables

Table 1. External data in prior literature.....	8
Table 2. Prior research on IT, IS, strategic sourcing, and data sourcing	9
Table 3. Perspectives on data sourcing in IS literature based on Jarvenpaa & Markus (2020).....	12
Table 4. Thesis structure: research streams and essays	18
Table 5. Data sourcing practices for sustainability	31

1 Introduction

In the age of digital transformation, data has become a valuable resource and the cornerstone for new business models, decision-making, and operational excellence (Buhl et al., 2013; H. Chen et al., 2012; Provost & Fawcett, 2013; Wixom & Ross, 2017). Empirical evidence proves that, businesses relying on data-driven decision-making have higher market values and profitability compared to those who did not (Brynjolfsson et al., 2011), have a significant impact on asset productivity in supply chain and business growth (D. Q. Chen et al., 2021), or have even revolutionized their respective industries (Parvinen et al., 2020). For the European Commission (2022a), data is “an essential resource for economic growth, competitiveness, innovation, job creation and societal progress.” The Commission specifically emphasizes the benefits of data originating outside companies, e.g., promoting cross-sectoral collaboration, knowledge sharing, and innovation, thereby encouraging businesses to share data that benefits society as a whole with other stakeholders (European Commission, 2020a, 2022a).

These benefits are also highlighted by analysts and consultancies – such as Forrester (Belissent, 2019), Deloitte (Schatsky et al., 2019), and McKinsey (Aaser & McElhaney, 2021) – suggesting that enterprises use third-party data (e.g., social media, weather, geospatial and satellite, web-harvested, or IoT data) to gain valuable insights and a competitive edge in their respective domains, although the use thereof is not yet widespread. Information Systems (IS) research occasionally provides evidence that companies are increasingly using data from external sources to improve advanced analytics, enrich business processes, decrease internal data curation efforts, and create new services (Baecke & Van den Poel, 2011; Baud et al., 2002; Strand & Syberfeldt, 2020). Even though the use of external data has been mentioned in the enterprise context since the late 1990s (Čas & Meier, 1999), we currently lack an academic understanding of the characteristics of external data and how enterprises should source and manage it.

While the benefits of sourcing external data in the enterprises are clear from a practitioner’s perspective, most studies assume that it is simply about getting the data, without specifying exactly how to achieve this. Jarvenpaa and Markus (2020) point toward a substantial but yet to be addressed void regarding data sourcing. In their first attempts to conceptualize the phenomenon, they refer to data sourcing as “procuring, licensing, and accessing the data” (Jarvenpaa & Markus, 2020).

This thesis aims to lay the foundation for data sourcing and, thereby, to address a real concern of enterprises that strive to actively source external data. Therefore, the overarching goal of this

thesis project is to advance data sourcing practices in the enterprise context. The sub-questions and contributions of this thesis are structured along three principal research streams.

The first research stream investigates the company-wide sourcing and managing of external data and proposes a taxonomy to inform data sourcing decisions, as well as the reference process for the sourcing and managing of external data. The second research stream, deep diving into one of the notorious external data types, discusses the extent to which open data is ready for use and suggests a method to screen, assess, and prepare open data for use. The third research stream tackles one of the most strategic topics facing enterprises – sustainability – by unveiling the underlying data sourcing practicing developed by companies in response to exerted institutional pressures, and by providing insights into four key initiatives in the field of environmental sustainability.

This introductory paper provides an overview of the thesis by presenting the three research streams, along with their motivations, research questions, and outcomes, respectively. Following the Introduction, the remainder of the paper is structured as follows: Section 2 presents the theoretical background of the thesis, introducing the external data and data sourcing literature, as well as the identified research opportunity. Section 3 details the overall research objectives of the thesis and its overarching structure, followed by a description of the research setting. Sections 4 to 6 present the research streams, outlining the individual motivations, research objectives, methodologies, main contributions, and discussions of each. Finally, section 7 provides an overview of our findings, critically discusses their overall implications and limitations, and outlines future research avenues in the field of data sourcing.

2 Background

2.1 External data as an under-researched topic

Since the late 1990s, the concept of external data has occasionally been mentioned in IS literature (Čas & Meier, 1999; Devlin, 1997). Delvin (1997, p. 135) defined external data as “business data (and its associated metadata) originating from one business that may be used as part of either the operational or the informational processes of another business.” To this day, this definition stands out among the few attempts to reveal and demystify what is concealed by the notion of external data in the enterprise context. Despite the increasing interest in and demand for external data in practice, the statement of Strand et al. (2003) still holds true: “It is not possible, in the literature, to find one common definition of external data.” Despite this lack of a singular, unambiguous definition, (IS) research nevertheless provides several examples of how companies use external data. These scenarios, as outlined in the literature (see Table 1), not only promote a better contextualization and understanding of enterprises’ motivation to use external data, but they also shed light on a large variety of external data sources.

Table 1 shows that companies source external data for various purposes. As highlighted by Jarvenpaa and Markus (2020): “... companies do not just source one type of data from one source.” When pursuing a competitive edge, companies use external data on markets, competitors, and the environment to improve their customer acquisition processes and to identify new, profitable customers (Strand & Carlsson, 2008). For these purposes, companies collect data from publicly available resources or acquire market reports or customer lists from third-party suppliers (Baecke & Van den Poel, 2011). For instance, Baud et al. (2002) refer to the use of external data on publicly-released losses to measure a company’s operational risks. More importantly, scholars regard external data as a boost to companies’ analytics potential, provided that a value-generating business context exists (Čas & Meier, 1999; Arndt & Gersten, 2001; Strand & Syberfeldt, 2020). Other scenarios mention master data management. For instance, Clevon and Wortman (2010) indirectly refer to external data when they discuss value-adding master data enrichment. Furthermore, data-driven business models can be built by leveraging external data (Sorescu, 2017).

Author	Context	Scenario	External data types and sources	Characteristics of external data
Čas & Meier (1999)	Marketing	Decision-making / analytics: "...preparing decisions concerning marketing activities of an enterprise, data are needed from internal sources..., as well as external data sources..."	Market studies, press releases from competitors	(not specified)
Arndt & Gersten (2001)	Direct marketing	Process improvement: "In many cases, purchasing additional data from outside the enterprise (external data) can enhance the overall data situation for the direct marketing tasks..."	Paid external data: business addresses Public information: world wide web, yellow pages	(not specified)
Baud et al. (2002)	Operational risk management	Process improvement: "internal data should be supplemented with external data in order to improve the accuracy of capital measurement."	Publicly released losses, databases based on a consortium of banks	Combination of external and internal data: "...external data may be viewed as 'implicit internal data,' meaning that external and internal data can be pooled together provided external data have been made comparable with internal data."
Strand et al. (2003)	Data warehousing	Decision-making / analytics: "as it has become more and more important to keep track of the competitive forces, it has become apparent that ...external information is (also) crucial."	Statistics institutes, syndicate data suppliers, industry organizations, the internet, industry sector data, business partner data, governmental/state data	External data definition based on Delvin (1997, p. 135): "Business data (and its associated metadata) originating from one business that may be used as part of either the operational or the informational processes of another business."
Strand & Carlsson (2008)	Decision support systems, business intelligence, data warehousing	Process improvement and decision-making / analytics: "External data is used in strategic, managerial and operational business and decision processes."	Acquired data: syndicate data suppliers, statistical institutes, industry organizations, county councils and municipalities, the Internet, business partners	Data quality: "The manual data quality controls are very costly." Data refinement: "In order to survive and sustain their competitive edges, the suppliers are spending a lot of resources on refining and enriching the raw data."
Cleven & Wortmann (2010).	Master data management	Master data enrichment: "...master data may further be enriched by adding organizational and/or technical metadata as well as external data in order to supply additional value."	(not specified)	(not specified)
Baecke & Van den Poel (2011)	Customer relationship management	Master data enrichment: "...companies constantly try to augment their database through data collection themselves, as well as through the acquisition of commercially available external data."	Commercially available data from external vendors: demographic, socio-economic, and lifestyle variables, related to a specific product category	(not specified)

Author	Context	Scenario	External data types and sources	Characteristics of external data
Piccoli & Pigni (2013)	Digital data streams, Big Data	Decision-making / analytics: "...external data streams from multiple sources ... combines to provide its customers with precise real-time traffic intelligence."	Public, business, individual, or community data: GSM and GPS probe data from external data streams, e.g., TomTom	Data streams characterized by the type of used technologies: application programming interfaces, web crawlers
Kwon et al (2014)	Big Data analytics	Decision-making / analytics: "processing external data for sense making becomes an integral part of big data analytics."	Public or commercial: customer information, market pressure, competitors, political regulations, and macroeconomics	Lack of control: "External data are obtained from sources over which a firm has little or no control."
Zhao et al. (2014)	Big Data analytics	Decision-making / analytics: "...data from external sources that will help generate new insights and provide competitive advantages."	Commercially available data from vendors: customer surveys, market intelligence, "internet-based sources such as social networking websites."	Price: "requires survey service expenses", "data capture is expensive and infrequent".
Zrenner et al. (2017)	Use of external data for supply network structures	(not specified)	Authors distinguish between two types of external data: open and closed (access restrictions, e.g., web services from a supplier or data provider)	Storing and maintenance outside of the internal systems / databases: "External data sources are not available in the company's IT infrastructure."
Sorescu (2017)	Data-driven business model innovation	New business models: "...companies can leverage internal and external data to generate new business models..."	Social media: API aggregation from specific providers (e.g., Twitter)	Extraction: "...extract only the portion that was of value to consumers."
Hopf (2019)	Predictive analytics	Decision-making / analytics: predictive business analytics	Two types of external data: published online or purchased from providers. Socio-demographic data, environmental data, public statistical data, geographic data, calendar events, website content, social media data	Storing and maintenance outside of the internal systems / databases and ownership: "This data stem neither from company IT systems, nor is the company owner of the data."
Strand & Syberfeldt (2020)	Business intelligence, decision support systems, analytics	Decision-making / analytics: "external data sources...are used jointly to allow for descriptive and predictive analytics, as well as prescriptive analytics."	Open data from public authorities: map data, road data, property data, civil data, traffic data, weather data	Storing and maintenance outside of the internal databases: "External data is any data stored or maintained outside the particular database of interest."

Table 1. External data in prior literature

The existing studies highlight the benefits of combining internal and external sources (Baud et al., 2002; Strand & Carlsson, 2008); an aspect sometimes referred to as "data augmentation" (Baecke & Van den Poel, 2011) or "enrichment" (Cleven & Wortmann, 2010). They also identify specific challenges, most importantly the lack of control (Kwon et al., 2014) and the transformation of external data to make it usable along with internal data (Baud et al., 2002; Strand & Carlsson, 2008). Accordingly, external data must be sourced and combined with

internal data (Baud et al., 2002), particularly in the context of Big Data and advanced analytics (Kwon et al., 2014; Zhao et al., 2014).

We conclude that prior literature on external data is fragmented and that it mostly justifies the use of external data, rather than elaborating on data sourcing.

2.2 Emergence of a data sourcing paradigm

Despite the increasing demand for external data, data sourcing has not been extensively discussed in the literature and is, instead, simply seen as getting the data. The success of sourcing decisions is a well-known concern of IT and IS sourcing (Kotlarsky et al., 2018), and we argue that this is also the case for data sourcing decisions. Studies on IT and IS sourcing have found that although cost reduction is a major factor in sourcing decisions, other factors, such as expertise, skills, quality improvement, and focusing on core capabilities, are becoming increasingly important (Clark et al., 1995; Könning et al., 2019; Lacity et al., 2010; Nevo & Kotlarsky, 2020; Oshri et al., 2015). In the context of data sourcing, because sourcing decisions are often viewed as routine (Jarvenpaa & Markus, 2020), a similar discussion is still found to be lacking. Table 2 provides an overview of these different but related sourcing types by comparing their respective objects of sourcing and definitions of IT, IS, strategic sourcing, and data sourcing.

	IT sourcing	IS sourcing	Strategic sourcing	Data sourcing
Definition	“...the delegation, through a contractual arrangement, of all or any part of the technical resources, human resources, and the management responsibilities associated with providing IT services to an external vendor” (Clark et al., 1995)	“...a broad umbrella term that refers to the contracting or delegating of IS- or IT-related work (e.g., an ongoing service or one-off project) to an internal or external entity (a supplier)” (Kotlarsky et al., 2018)	“...a critical area of strategic management that is centered on decision-making regarding an organization’s procurement activities such as spend analysis, capability sourcing, supplier selection and evaluation, contract management and relationship management” (Rafati & Poels, 2015)	“...procuring, licensing, and accessing data (e.g., an ongoing service or one-off project) from an internal or external entity (supplier)” (Jarvenpaa & Markus, 2020)
Object of sourcing	Hardware, software, and related services which meet the specific technology needs of the organization	IS- or IT-related services requiring specific expertise, which may include hardware and software, as well as infrastructure to manage and store information	Raw materials and components, finished goods, services, equipment and machinery, and staffing	Data and data-related services

Table 2. Prior research on IT, IS, strategic sourcing, and data sourcing

Jarvenpaa and Markus (2020) refer to data sourcing as “procuring, licensing, and accessing data (e.g., an ongoing service or one-off project) from an internal or external entity (supplier).” This definition builds upon the concept of IS sourcing, which implies contracting or delegating IS- or

IT-related work (Kotlarsky et al., 2018). It is worth noting that recent agro-geoinformatics literature (Sun et al., 2021) discussed different forms of data sourcing. First, conventional data sourcing, which refers to obtaining data from a variety of sources and which typically involves the finding, obtaining/purchasing, assessing, integrating, and use of data. Second, crowd-based data sourcing, which emerges as a “data procurement paradigm that engages Web users to collectively contribute and process information” (Amsterdamer & Milo, 2015). Third, cloud-based data sourcing, which implies that the cloud-stored data is accessed via dedicated platforms such as Amazon Web Services and Microsoft Azure (Sun et al., 2021). Despite its relevance, conventional data sourcing, which is at the center of Sun et al.’s (2021) study, only plays a minor role in the scarce IS literature on the topic. It is only briefly mentioned in specific scenarios, e.g., data warehouse enhancement with external data (Strand & Carlsson, 2008), or occasionally mentioned in consulting reports (Aaser & McElhaney, 2021; Schatsky et al., 2019).

Data sourcing entails multiple challenges, among others, the sources’ variety and complexity (both external and internal sources), data quality issues, legal and regulatory considerations, and the general role of data in business strategy (Jarvenpaa & Markus, 2020). Apart from the obstacles related to the data as such, data sourcing requires the establishment of inter- and intra-organizational relationships, frequently seen as customer-supplier relationships (Jarvenpaa & Markus, 2020). From this perspective, data sourcing resembles strategic sourcing; while the former focuses on data acquisition, the latter generally involves the acquisition of goods and services. In their first attempt to structure the field, Jarvenpaa and Markus (2020) – relying on typical sourcing perspectives – distinguish between the following three perspectives (see Table 3):

1. The commodity or transactional perspective, which emphasizes the role of data as a commodity that can be bought and sold in the market. The core idea behind this view is that data has value on its own (regardless of its use context) and can be treated as a tradable asset like oil or gold. In line with transaction cost theory, this perspective, which is a common perspective in IS sourcing literature (Lacity et al., 2016), assumes that data is a homogeneous resource that can be easily compared and evaluated. The data is assumed to be easily harvestable for the creation of value-added services (Piccoli & Pigni, 2013), e.g., via open data platforms. Therefore, transaction costs become the basis for data sourcing decisions and strategies (Jarvenpaa & Markus, 2020). Generally, the transactional perspective on data sourcing is helpful to understand the economic aspects, but it may not capture the full complexity of the data sourcing process.

2. The relational perspective, which enhances the prominence of the inter-organizational context of data sourcing and the role of external relationships. This perspective implies that data sourcing is a collaborative process that requires the development of strong partnerships between organizations. It views external data sourcing as a strategic activity that requires trust and collaboration between the enterprise and external parties, where data, as a strategic asset, can be leveraged for competitive advantage through the creation of mutually beneficial partnerships. It also acknowledges that data can be sourced through a variety of organizational arrangements, ranging from bilateral relationships with data providers to multilateral relationships in which data is shared or exchanged with peers, which require agreeing on technical data specifications and the related conditions. According to this view, data is assumed to travel across different use contexts and to take on different forms, e.g., the use of restricted health data in the for-profit corporate environment (Winter & Davidson, 2019).
3. The processual perspective emphasizes “the value of entanglement of data and operations on data that could take place at any point, from the source to the final reuse” (Jarvenpaa & Markus, 2020). It accentuates the intra-organizational context of data sourcing and starts from the assumption that the sourcing decision and organizational arrangement are control-based rather than cost-based. According to this perspective, data sourcing is a complex process that involves several distinct steps and activities, including identifying data needs and goals, identifying potential data sources, evaluating and selecting data sources, and acquiring and integrating data so that it can be used.

Perspective	Description	Data characteristics	State of IS research and exemplary topics
Commodity / Transactional	Emphasizes that transaction costs are crucial when making the external data sourcing decision: e.g., the cost of acquiring the data, combination efforts, penalties for licensing violations, and access restrictions.	Considers data as a homogeneous resource, which is easily harvestable and offered “as-is”.	Prevails in the IS literature (Y. Chen et al., 2017) and mentioned in the contexts of databases, software programs, data traces, data records and information artifacts (Abraham et al., 2019; Pigni et al., 2016; Tallon, 2013).
Relational	Focuses on the inter-organizational context, which enables data sourcing, and the organizational arrangements based on trusted relationships.	The data is assumed to travel across different use contexts, e.g., in inter-organizational data exchanges (Winter & Davidson, 2019). Relationship and ownership are important value-adding factors for repurposed data (Jarvenpaa & Markus, 2020).	Almost non-existent in IS literature, and only discussed in the context of big data governance (Winter & Davidson, 2019) and research on data communities (Leonelli, 2015).

Perspective	Description	Data characteristics	State of IS research and exemplary topics
Processual	<p>Focuses on the intra-organizational context, where data is “temporal, co-dependent, indeterminant, and pervasively editable” (Jarvenpaa & Markus, 2020).</p> <p>It is important to recognize the interconnection between data and the actions performed on it throughout its lifecycle (Jarvenpaa & Markus, 2020).</p>	In terms of this perspective, data cannot provide any information until it is sought, chosen, extracted, and interpreted, and any purported data that cannot provide any information is not considered as data (Jones, 2019).	Emerging in IS literature, primarily reflected in the contexts of electronic medical records (Jones, 2019; Wadmann et al., 2013), social media (Orlikowski & Scott, 2014), and digital platforms (Aaltonen & Tempini, 2014)

Table 3. Perspectives on data sourcing in IS literature based on Jarvenpaa & Markus (2020)

2.3 Research opportunity

Although companies are increasingly using external data, their underlying usage processes lack dedicated methodological guidance. From their review of the scarce literature, Jarvenpaa and Markus (2020) conclude that while organizations source external data for different purposes, information systems research (especially on data governance and data quality) lacks a sourcing perspective. While the transactional view dominates the perspectives on data sourcing, processual and relational views receive little attention. We, in turn, argue that all three perspectives are important in the enterprise context and, therefore, should be considered when sourcing data.

3 Thesis overview

3.1 Research objectives

With regard to the identified research opportunities, this thesis project aims to **advance data sourcing practices in the enterprise context**. As a comprehensive approach to this problem, it brings together three research streams (see Figure 1). After considering external data sourcing and clarifying its foundations in stream 1, we deep-dive into one of the most prominent external data types (i.e., open data) in stream 2 and analyze data sourcing practices regarding sustainability in stream 3, being a crucial strategic topic that currently confronts enterprises.

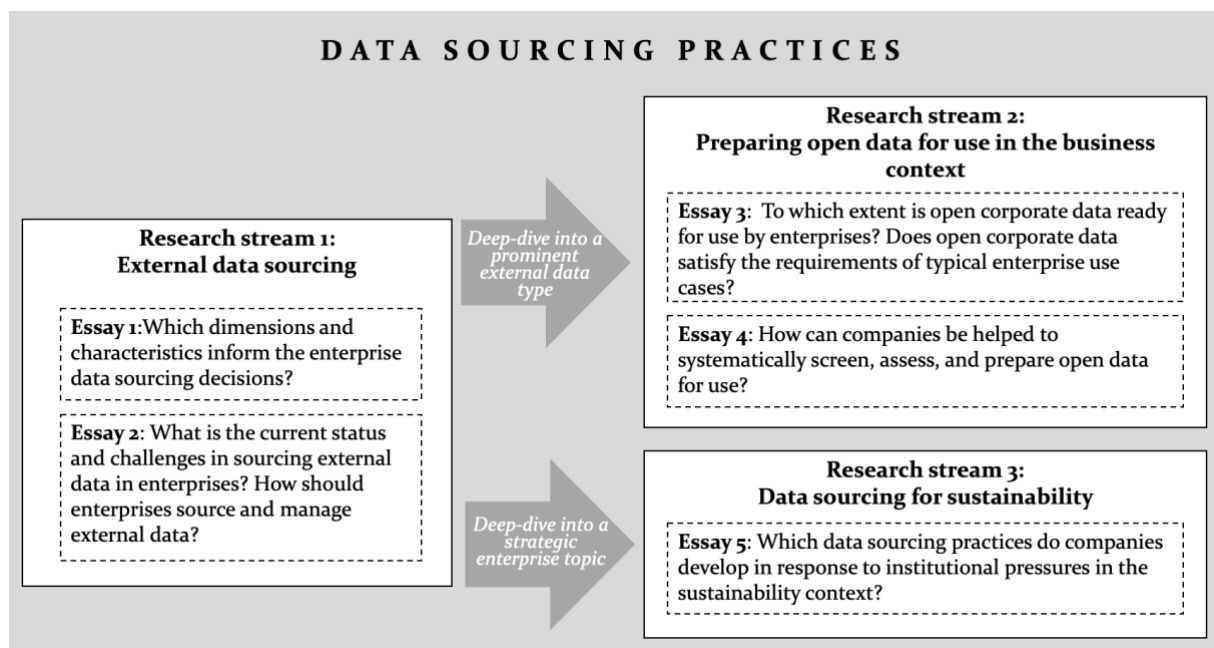


Figure 1. Overview of research streams and questions

To begin with, the first research stream lays the foundations for external data sourcing by clarifying the object of sourcing and the relevant factors influencing sourcing decisions, as well as the sourcing processes and the underlying design principles of the sourcing and managing of external data. Given its increasing importance in an enterprise setting and its multifaceted nature, this stream addresses data sourcing in two essays. Essay 1 informs the data sourcing decisions by proposing a data sourcing taxonomy covering the three perspectives, namely the transactional, relational, and processual perspectives (see subsection 2.2). Essay 2 proposes a reference process for the sourcing and managing of external data, accompanied by design principles – methodologically grounded in design science research – that guide enterprises through uncertain steps and unfamiliar territory.

The second research stream deep-dives into open data use, as one of the commonly used types of external data (Hopf, 2019; Roeder et al., 2020; Strand & Syberfeldt, 2020). This stream challenges the common perception that open data is a readily accessible and easily usable type of external data. The goals of this stream are, among others, to understand open data's readiness for use in the enterprise setting and to assist companies in using open data. Essay 3, going beyond existing approaches, conducts a use case-driven assessment of open corporate registers and considers three distinct levels: metadata, schema, and content. The outcome of this study is the result of our ready-for-use assessment of 30 corporate registers in four concrete use cases. Essay 4 integrates the findings of Essay 3 and addresses the lack of perspectives on how to efficiently prepare open data for productive use. Its main contribution results in a four-phased method to screen, assess, and prepare open data for use, thereby supporting companies in their open data activities.

The third research stream confronts sustainability, being one of the most compelling topics of present-day companies. Reporting on sustainability goals requires the collection, processing, and interpreting of large amounts of data (e.g., related to emissions or recycled materials), which were neither captured nor analyzed previously. Therefore, the objective of this research stream is to better understand the existing data sourcing practices in enterprises for selected sustainability scenarios. Essay 5's main contributions are twofold: First, as theoretical contribution, we propose a framework based on institutional theory to explain how companies develop their data sourcing practices in response to regulatory, normative, and cultural-cognitive pressures. Second, our empirical contributions include insights into five case studies that represent key initiatives in the field of environmental sustainability, that touch on first, understanding the ecological footprint, and second, obtaining labels or complying with regulations, both on product and packaging levels.

In line with the overarching research goal, Table 4 provides a detailed overview of the defined research streams and their related essays, respectively outlining the main research questions, methods used, and key contributions per essay, as well as the current publication status of each essay.

3.2 Research setting

Our research was carried out in the context of the Competence Center Corporate Data Quality (CC CDQ). The CC CDQ is a consortium research project (Österle & Otto, 2010) that assembles data management experts from approximately 20 multinational companies (the exact number varies annually), including a team of researchers. In the tradition of collaborative practice

research, consortium research promotes exchanges between researchers and practitioners by focalizing their common interest and “serving the general knowledge interest as well as knowledge interests that are specific for the participating organizations” (Mathiassen, 2002).

The CC CDQ consists primarily of large multi-national companies representing various industry sectors, e.g., retail, fast-moving consumer goods, automotive, chemical engineering, and pharmaceutical. Obviously, their respective data management initiatives have different levels of maturity, as well as varying goals and challenges, allowing for a diversified experience exchange and insights. This setting, in particular, is welcomed as it not only increases an understanding of the data sourcing approaches within the respective companies but also addresses their relevant practical problems with scientific rigor. To gather additional insights and enrich our results, we also reached out to an extended network of practitioners and academic experts beyond the confines of the original research consortium. This contributed to a more robust validation of our research findings and increased the generalizability of our results.

The consortium research setting provides a favorable environment for design-oriented research (Legner et al., 2020; Vom Brocke & Buddendick, 2006). This is particularly relevant since the research goal of this thesis project is to advance data sourcing practices by going beyond a mere observation or analysis of the phenomenon of interest. In respect of each of the three research streams, our research activities were embedded in larger projects aimed at addressing challenges which companies face with regard to external data sourcing. Our research activities were conducted in accordance with the nominal steps of the consortium research methodology (see Figure 2), namely analysis, design, evaluation, and diffusion, which are described in more detail in the related essays. The consortium research allows the leveraging of implicit practitioner knowledge and thereby informs our own research in the emerging research field.

In line with the analysis phase of consortium research (see Figure 2), we rely on qualitative research methods to understand the requirements, motivations, and challenges of the CC CDQ member companies and to leverage their practical knowledge. Within all research streams we used focus groups and plenary discussions, mainly for the purpose of garnering insights into our research topics. For instance, in streams 1 and 3 we conducted semi-structured expert interviews with subject matter experts to better grasp the complexity of the topic and to collect company-specific insights. We also relied on the focus groups and plenary discussions to inform the design of our artifacts, and to collect and evaluate feedback on these artifacts. This aligns with the guidelines for the third phase of consortium research, i.e., evaluation, allowing us to evaluate our results against the research objectives (Österle & Otto, 2010). Throughout the whole thesis,

traditional research activities, such as a literature review and desk research, were conducted to understand the existing body of knowledge and to inform the design of the artifacts and their research implications.

In the context of this thesis, consortium research allows conducting longitudinal design science research in multilateral settings (Legner et al., 2020), while providing an umbrella for detailed research activities. In line with the abovementioned research goals (see subsection 3.1), each essay helps to analyze existing data sourcing practices and further develop them in close research-practice interactions, using different methodological configurations. In the first research stream, we focused on designing a data sourcing taxonomy (Essay 1) by adhering to the guidelines proposed by Nickerson et al. (2013). In Essay 2, we employed the design science research (DSR) methodology following Peffers et al.'s (2007) model to develop a reference process, using an objective-centered solution as a research entry point. This entailed sequential design and development phases, followed by subsequent demonstration and evaluation in company-specific instantiation (naturalistic evaluation), and also including focus groups and experts. In the second research stream, we started with the exploration of open corporate datasets (Essay 3) from the consumer perspective. In order to develop a method to screen, assess, and prepare open data for use in the enterprise context, Essay 4 adopts action design research (ADR) and is driven by the insights gained from practical implementations (Sein et al., 2011), i.e., productive platform for data quality services and prototype of an open data catalog. In contrast to the method used in Essay 2, that relegates demonstration and evaluation to a subsequent phase, ADR incorporates evaluation into the design cycles (Sein et al., 2011). In doing so, our design science-oriented essays employ a combination of systematic artifact development, comprehensive demonstration, and rigorous evaluation. Essay 5 leverages the consortium research setting for the analysis of cases and sets the problem space for future design science research activities.

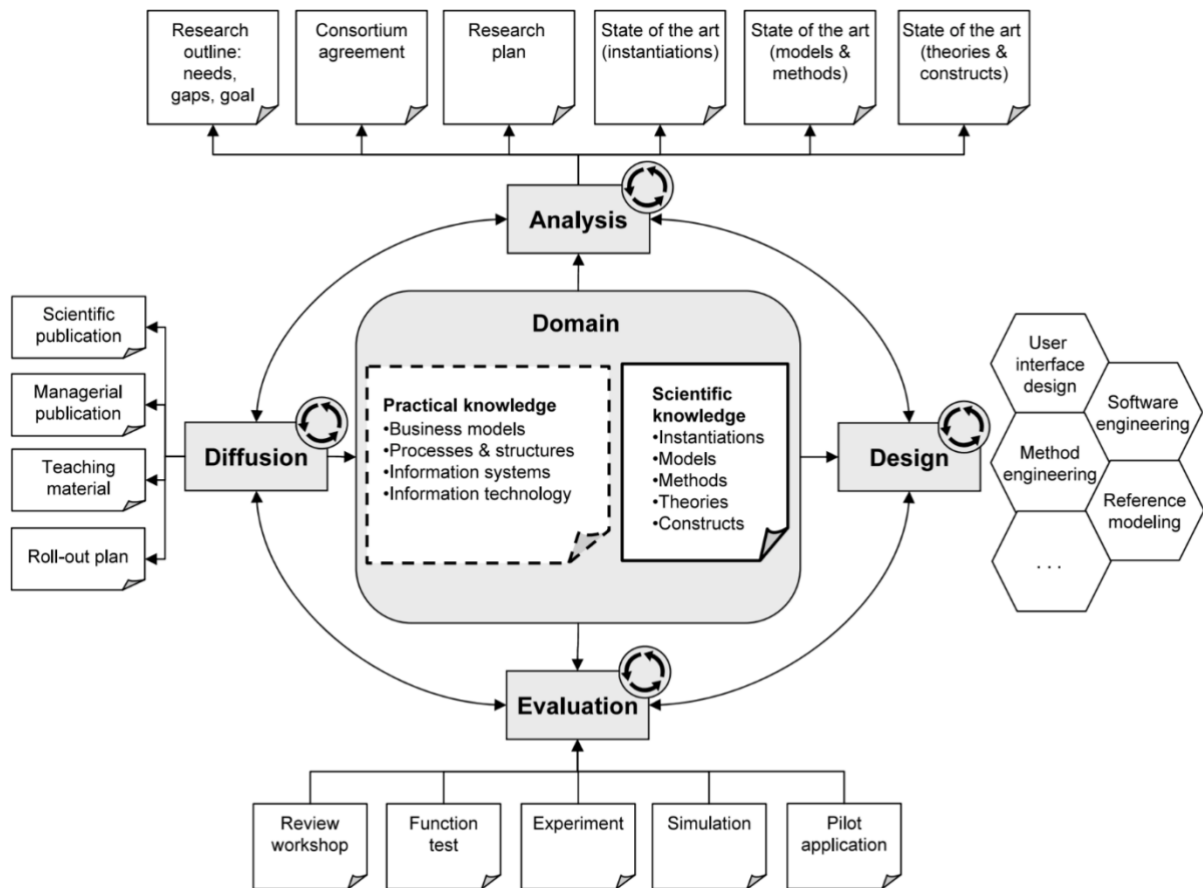


Figure 2. Consortium research overview (Österle & Otto, 2010)

Essay	Research question(s)	Methodology	Key contributions	Publication status
Research stream 1: External data sourcing				
Essay 1: Responding to the Siren Song of External Data: A Taxonomical Approach to Data Sourcing	Which dimensions and characteristics inform the enterprise data sourcing decisions?	Taxonomy development (Nickerson et al., 2013), with conceptual-to-empirical and empirical-to-conceptual iterations	Data sourcing taxonomy and classification of data sources used by practitioners in real-world use cases	<i>First version:</i> presented at pre-ICIS 2021 SIG Advances in Sourcing <i>Extended version:</i> submitted to the Electronic Markets – The International Journal on Networked Business
Essay 2: Unleashing the Potential of External Data: A DSR-based Approach to Data Sourcing	What is the current status and challenges in sourcing external data in enterprises? How should enterprises source and manage external data?	Design science research (Peppers et al., 2007)	Reference process for sourcing and managing external data	Published in the proceedings of the 30th European Conference on Information Systems (2022)
Research stream 2: Preparing open data for use in the business context				
Essay 3: Open Data in the Enterprise Context: Assessing Open Corporate Data's Readiness for Use	To which extent is open corporate data ready for use by enterprises? Does open corporate data satisfy the requirements of typical enterprise use cases?	Literature review, focus groups with practitioners, in-depth analysis of the open datasets	Use case-driven analysis of open corporate registers	<i>First version:</i> published in the proceedings of the 9th International Conference on Data Science, Technology and Applications (2020) <i>Extended version:</i> published in Communications in Computer and Information Science book series, volume 1446 (2021)
Essay 4: A Method to Screen, Assess, and Prepare Open Data for Use	How can companies be helped to systematically screen, assess, and prepare open data for use?	Action design research (Sein et al., 2011)	Method to screen, assess, and prepare open data for use	<i>First version:</i> published in the proceedings of the 16 th International Conference on Design Science Research in Information Systems and Technology (2021) <i>Extended version:</i> accepted to the special issue of the Journal of Data and Information Quality on Quality Aspects of Data Preparation
Research stream 3: Data sourcing for sustainability				
Essay 5: Introducing a Data Perspective to Sustainability: How Companies Develop Data Sourcing Practices for Sustainability Initiatives	Which data sourcing practices do companies develop in response to institutional pressures in the sustainability context?	Explorative research design based on multiple case studies (Benbasat et al., 1987; Yin, 2009)	Research framework based on institutional theory, instantiated with four sustainability initiatives	<i>First version:</i> presented at pre-ICIS 2022 SIGGreen <i>Extended version:</i> accepted to the special issue of the Communications of the Association for Information Systems on Digital Innovation for Social Development and Environmental Action

Table 4. Thesis structure: research streams and essays

4 Research stream 1: External data sourcing

4.1 Motivation and background

“With our own data we can only look internally. We need to see industry benchmarks, regional trends, what waves we can ride on; we derive competitive advantage by getting data from outside and enhancing our own data” (Chief Data Officer of Flagstar Bank, as cited in Belissent, 2019). An increasing number of studies (see subsection 2.1) demonstrate that combining internal data with external data improves advanced analytics, enriches business processes, decreases internal data curation efforts, and creates new services (Baecke & Van den Poel, 2011; Baud et al., 2002; Schatsky et al., 2019; Strand & Syberfeldt, 2020). Despite the increasing interest in the topic, external data continues to be an underexploited resource: Finding and accessing suitable external data sources, unclear ownership, uncertainty of the data quality, and the costs of sourcing data are some of the reasons why the use of external data in an enterprise context has not yet become commonplace (Van Alstyne et al., 1995; Arndt & Gersten, 2001; Janssen et al., 2012). Based on their review of scarce literature, Jarvenpaa and Markus (2020) conclude that while organizations source various types of data for different purposes, information system research, data governance, and data quality are devoid of a sourcing perspective.

In line with Jarvenpaa and Markus (2020), we argue that – given the increasing demand for external data – companies should develop a more professional approach to data sourcing. Similar to IS and IT sourcing (see subsection 2.2), sourcing decisions on external data should consider different dimensions, among others, the technical specifications and costs of data acquisition, as well as the relationships among and dependency on the parties that provide the data (Jarvenpaa & Markus, 2020). While some attempts have been made to facilitate the selection of datasets in specific contexts (Kruse et al., 2021), a data sourcing decision extends beyond the mere selection of suitable datasets (as the object of sourcing) and also requires contractual structures and the management of the entire process, from finding and obtaining data to integrating and using it.

4.2 Research objectives, methodology, and contributions

The goal of the first research stream is twofold: First, we seek to clarify the object of data sourcing and the relevant factors that inform data sourcing decisions. Second, we aim to elaborate on data sourcing practices and develop a reference process for the sourcing and managing of external data. Therefore, we undertake this endeavor through two essays.

To begin with, Essay 1 aims at answering the research question: “*Which dimensions and characteristics inform enterprise data sourcing decisions?*” We develop a data sourcing taxonomy to answer this question, based on Nickerson et al.’s (2013) guidelines. With little data available on the objects and significant understandings of the domain, we begin with two conceptual-to-empirical iterations, followed by another empirical-to-conceptual iteration to build a meaningful taxonomy. In doing so, the resulting taxonomy combines theoretical knowledge and empirical findings. It outlines eight relevant dimensions and 29 related characteristics to cover the three perspectives on data sourcing decisions proposed in the literature. From a transactional perspective our findings highlight access conditions, licensing, and price; from a relational perspective the emphasis is on contractual parties and data ownership; and from a processual perspective the focus is on data access, data preprocessing, and data use. For the same scenario, we illustrated the enterprise use of the taxonomy with three sourcing options, in which practitioners employ three different external datasets to support their master data activities, reducing the data management efforts for business partner data. Beyond the importance of considering all three perspectives for an informed data sourcing decision, we provide a deeper understanding of their specific characteristics and unique considerations. For instance, the nature of relationships with contractual parties can have a significant impact on the success of a sourcing decision as it influences the quality of the data received and the level of support provided during onboarding and consequent management processes. Additionally, from a technological point of view, the access to the data can also impact on the effectiveness and efficiency of data usage. Therefore, it is crucial to consider all the abovementioned characteristics holistically when making a sourcing decision, and to prioritize those that are most important to the intended use case and business objectives.

Essay 2 aims at answering the following research questions: “*What is the current status and challenges in sourcing external data in enterprises?*” and “*How should enterprises source and manage external data?*” Embedded in industry–research collaboration, we adhere to design science research by following Peffers et al.’s (2007) methodology and seek to design actionable guidance for sourcing and managing of external data in a form of a reference process. As a particular instantiation of a reference model and a generic procedure for evidence-based IS research (Goeken, 2011), a reference process aims to generalize the usual process sequence and its elements, such as activities and milestones (Becker et al., 2007; Wilmsen et al., 2020). Reference models, as a specific type of conceptual model (Frank, 2014; Vom Brocke, 2007), are widely employed in research and industry for designing complex systems, facilitating communication with users, and forming a solid basis for implementation (Frank, 1999). They

expedite the development of enterprise-specific models (Fettke & Loos, 2003), aligning perfectly with our research objectives. In the realm of data management, reference models have proven to be useful artifacts that accumulate design knowledge from academic and practitioner communities (Legner et al., 2020). They are constructed by modelers, who describe the universal elements and relationships of a system, providing recommendations, “thus creating a center of reference” (Ahlemann & Riempp, 2008, p. 89). In our case, the reference nature of our artifact emerged from analyzing common practices and jointly designing best practices with practitioners, driven by the innovation stimuli (Becker et al., 2002). Following the design science principles, reference models undergo iterative design and evaluation processes (R. Winter & Schelp, 2006). We followed the six steps of the DSR process model (Peffer et al., 2007), we conducted two distinct design cycles, resulting in a reference process that supports enterprise-wide data sourcing activities.

Based on the scarce data sourcing literature and the empirical evidence from the design cycles, we identified six core phases of the process: start, screen, assess, integrate, manage and use, and retire. Each process step contains a clear input, a set of underlying activities, as well as related roles and techniques. It ends at a defined milestone, allowing a progress review along the reference process. The sequence of the phases is nominal, allowing for the simultaneous execution of activities if the necessary conditions are met. In line with our research goal, our findings enable a shift from ad-hoc sourcing practices to a well-defined approach for the sourcing and managing of external data.

4.3 Discussion, limitations, and outlook

Our work establishes a foundation for scientific inquiry about external data sourcing and demonstrates that this phenomenon is indeed more complex than “getting the data”. Our main contributions in this research stream (i.e., an external data sourcing taxonomy and a reference process for sourcing and managing external data) complement each other in their overarching goal of advancing external data sourcing practices.

From an academic perspective, both Essays 1 and 2 are among the first systematic and methodologically rigorous attempts to conceptualize data sourcing and advance data sourcing practices. The findings of this research stream provide the necessary toolbox for the investigation of future data sourcing paradigms, from the perspective of informed sourcing decisions (Essay 1) and the underlying processes (Essay 2). The proposed taxonomy (Essay 1) allows its users to position sourcing options for external data more clearly and it may inspire researchers to elaborate on these options in more detail. In addition, it helps to define and

analyze sourcing strategies and processes concerning external data. Our study contends that the understanding of sourcing decisions is a complex endeavor requiring an analysis of all options through the prism of transactional, relational, and processual perspectives. With regard to the reference process for sourcing and managing external data (Essay 2), our findings synthesize and expand the scarce body of knowledge on data sourcing by, specifically, extending the processual perspective (see subsection 2.2) with empirical evidence. We notice that the proposed reference process shows commonalities with strategic sourcing processes, but that it also uncovers data sourcing specificities. The latter includes, for instance, the importance of semantics and concept mapping to integrate external data. Among the major advantages of our reference process is its ability to guide enterprises in their sourcing activities in a systematic way with clear milestones.

From a practitioners' perspective, our findings were deemed useful to support enterprises in their external data sourcing activities. The taxonomy provides guidance to analyze external data when it is explored or used by a firm, as well as to determine the rationale behind the choice of the data source. The designed reference process, which aims to solve the increasingly relevant organizational challenges, contributes to the professionalization of external data sourcing. For instance, it contains the milestones which present well-defined decision points that allow the involved parties to effectively communicate, based on standardized performance criteria. Furthermore, to overcome the shortcomings of linear phase models, our reference process offers flexibility and process variability by formulating 11 possible variations of how the activities in the phases can be executed. They are also positioned as entry points, considering the current situation of the sourcing activities within the company.

In addition, our findings also contribute to an ongoing discussion on the definition of external data types, their specific properties, as well as the hurdles related to their practical use. Our reference process addresses the challenges associated with external data quality. Unexplored external datasets require a more thorough assessment than those of traditional quality metrics (R. Zhang et al., 2019). Since external data creators and publishers are detached from their users (i.e., enterprises), the latter have limited knowledge about the data's characteristics and underlying quality. In the case of repurposed data, it is essential to adopt an approach that provides multiple perspectives on the sourced data. Furthermore, contemporary literature does not provide a comprehensive overview of existing external data types, especially since the current definitions vary considerably, even for seemingly well-known types such as open data, social media and online sources, and sensor and IoT data (Kitchin, 2014). Aspects such as accessibility, machine-processability, and licensing have been long known to be foundational principles of open data (Open Government Working Group, 2007). Our external data sourcing

taxonomy proposes additional dimensions which could serve as a basis to describe the different facets of emerging external data types.

Although our study is novel and advances the topic of data sourcing in IS literature, it has certain limitations. In line with Nickerson et al. (2013), a taxonomy is useful in a best case but never perfect. While its suggested dimensions and characteristics are grounded in scarce literature and extensive empirical evidence, our taxonomy (as taxonomies in general) can be extended with additional content by including new discoveries on the use context of external data. In this sense, our proposed taxonomy is a first step toward a more comprehensive taxonomy for external data sourcing. Only its widespread application, both in academia and in practice, will reveal the extent to which the taxonomy is complete, or which other dimensions should be added. With regard to our reference process, although our findings were well-perceived throughout the instantiations and the focus group discussions, large-scale demonstrations or evaluations have not yet been conducted. Thus, we foresee future research activities to apply the reference process in diverse use cases and enterprise contexts, which would help generalize our findings and identify situational configurations. While our study focused on a reference process, it provides some first insights into emerging roles in the context of external data sourcing and management. Studying both aspects would allow us to develop a broader perspective on external data governance mechanisms. Another promising avenue for future research is the opportunity to explore and discuss the peculiar nature of digital data as a semantic resource, drawing upon emerging literature on the topic (Aaltonen et al., 2021).

5 Research stream 2: Preparing open data for use in the business context

5.1 Motivation and background

Open data is the most discussed external data type in prior literature (Hopf, 2019; Roeder et al., 2020; Strand & Syberfeldt, 2020). Open data can be defined as “data that is freely available, and can be used as well as republished by everyone without restrictions from copyright or patents” (Braunschweig et al., 2012). It offers business and innovation potential for companies (Janssen et al., 2012; Zuiderwijk et al., 2015) and national economies, estimated at a total open data market size of between 199.51 and 334.21 billion euros in the European Union by 2025 (European Commission, 2020c). However, as simple as the easy and free availability of open data may appear, open data consumers must overcome significant barriers, with the result that the actual use of open data remains well below expectations. Many of these obstacles are associated with data quality issues, e.g., a lack of transparency about its content, incomplete or missing data, or unclear licensing and access conditions (Bachtiar et al., 2020; Vetrò et al., 2016). These barriers hinder companies from leveraging open data’s value generating potential (Enders et al., 2020) and lead to a “mismatch between the needs and expectations of the users and the possibilities offered by available datasets” (Ruijter et al., 2018).

Despite rising expectations regarding open data use (Zuiderwijk et al., 2012), uncertain open data quality is continuously mentioned as one of the most prominent barriers to widespread open data adoption in an enterprise setting (Corsar & Edwards, 2017). To address this issue, researchers have developed dedicated assessment techniques, such as the “Luzzu” framework (Debattista et al., 2016), the “LANG” approach (R. Zhang et al., 2019), and the “QUIN” usability criteria (Osagie et al., 2017). Unfortunately, the assessment scope of these techniques is limited as they mainly consider the metadata level. Moreover, these techniques are not embedded in the complete set of activities required for enterprise-based open data consumption. For instance, they are poorly linked to data preparation, which includes techniques such as data collection, data integration, data transformation, and data cleaning (S. Zhang et al., 2003). To the best of our knowledge, suitable processes and methodological approaches that help prepare open data for enterprise use do not yet exist, at least not in a well-structured, holistic, and rigorous scientific manner. It therefore remains uncertain which process steps and actions qualify to identify, assess, and prepare open data for use, successfully.

5.2 Research objectives, methodology, and contributions

The overall objective of this research stream is to understand the open data's readiness for use in the enterprise setting and to propose how companies can be assisted in using open data. The collected insights pertain to the research questions that we posed in Essays 3 and 4.

Essay 3 addresses a set of two research questions: *“To which extent is open corporate data ready for use by enterprises?”* and *“Does open corporate data satisfy the requirements of typical enterprise use cases?”* To answer these questions, we conducted our research by adopting the following methodological approach: a literature analysis to understand open data's current state and its adoption barriers; focus groups with practitioners to specify use cases in the enterprise context; and an in-depth assessment of open corporate datasets in the form of metadata, schema, and content analysis. The focus group activities resulted in four concrete use cases and corresponding business concepts, all of which could potentially be sourced from open corporate datasets. Based on the input from practitioners and related literature, we considered 30 corporate registers in our analysis, which are provided by official government agencies that make their data available in full open access. We rigorously assessed the metadata of the identified registers to determine whether the desired data is usable or not, followed by a schema analysis to find common attributes between the registers, and concluded with a “ready for use” assessment to ascertain if the necessary attributes are present to satisfy the use case requirements. The main contribution of this first study resides in the results of our use case-driven analysis, which reveal that open corporate datasets are of limited use in typical use cases. Beyond the assessment of open corporate data, our study provides a methodological contribution by proposing a use case-driven approach comprising four steps: (1) the identification of the open data sources, (2) a metadata analysis, (3) a schema analysis of the datasets, and (4) a “ready for use” assessment based on a comparison of relevant business concepts in the selected use cases.

Essay 4 builds on and extends these findings. It aims at answering the question *“How can companies be helped to systematically screen, assess, and prepare open data for use?”* Therefore, in order to accumulate prescriptive knowledge with the due scientific rigor in an iterative research process, we adhere to the methodological stages suggested by Sein et al. (2011). We leverage on DSR to answer our research question, which builds on intentional, intellectual, and creative problem-solving activities (Chandrasekaran, 1990) for the “systematic creation of knowledge about, and with (artificial) design” (Venable et al., 2016). Building on the problem framing and theoretical foundations, we conducted two building, intervention, and evaluation (BIE) stages, focusing on artifact design. In the first BIE cycle was a part of a multiyear research

project that resulted in a productive platform for data quality services, operated by the data service provider. The first version comprises the method's nominal steps and the supporting use of knowledge graphs to explicate business concepts and link them to related datasets. It was evaluated with practitioners during five focus group discussions. The second BIE cycle was a two-year research project that aimed to build an open data catalog for business purposes and resulted in a prototype implementation. As part of our concurrent evaluation, we applied the method to more than 10 business scenarios (e.g., customs clearance, marketing, and customer analytics) to identify 40 open data use cases, screen and assess relevant open datasets, and map their data models.

Our research produced prescriptive knowledge in the form of a meaningful method to screen, assess, and prepare open data for use in an enterprise setting. The method comprises four phases and supports companies in all steps from deciding on the suitable use cases for open data to preparing them for actual use. It includes techniques and documentation templates (when appropriate) for the introduced steps. Our proposed method ensures a purposeful discovery and selection of open data sources and datasets, with consideration of relevant aspects such as provenance, licensing, and access conditions. It integrates a systematic approach to quality assessment of open datasets, being a major criterion for their selection and preparation for further use.

5.3 Discussion, limitations, and outlook

Our findings regarding research stream 2 validate the assumption that existing open data assessment methods require amendments to integrate the use context. To the best of our knowledge, this is one of the first systematic attempts to address the widespread sociotechnical barriers to and challenges of open data adoption (i.e., lack of transparency, heterogeneity, and the unknown quality of open datasets). Compared to prior literature, our findings consolidate different streams of open data research through a systematic approach: First, we contextualize open data use by providing guidance for use case documentation and by exemplifying the generic business scenarios which allow the user to gain value from open data. We thereby ensure that open data is “usable for the intended purpose of the user” (Welle Donker & Van Loenen, 2017). Second, building on our findings of Essay 3, our method (Essay 4) suggests a context-aware open data assessment approach that comprises metadata, schema, and content level techniques. It thereby reflects open data quality assessment approaches and links them to traditional data quality literature. Third, our method is enabled by using semantic concepts for data integration – a knowledge graph and reference ontologies – that allow mapping open datasets to internal

data objects. This approach enables enterprises to locate open datasets containing attributes that correspond to business concepts, which then relate to their internal data. It provides a scalable approach to the integration of heterogeneous datasets (Zuiderwijk et al., 2015; Bizer et al., 2009; Auer et al., 2007; Zaveri et al., 2016).

To assess the usability of open data in general, future research should place more emphasis on domain-specific and use case-specific analyses to complement our suggested methods. From a theoretical perspective, the concept of open data quality should be revisited with regard to usability (Bicevskis et al., 2018; Vetrò et al., 2016). A limitation of Essay 3 is that our analysis focuses on selected registers in countries that are deemed advanced with regard to open data provision (European Commission, 2020b). Given the total number of existing business registers, our sample does not allow us to draw conclusions about the entire domain. In addition, our assessment relies on four use cases identified by the focus group, but other potential use cases could be discovered. We also underline the need for domain ontologies, such as the euBusinessGraph (2019) common semantic model for company data, which could serve as a basis to provide more consistent and compatible open datasets across different open data portals and providers.

Since the method (Essay 4) comprises context-specific elements, it could also benefit from pre-existing reference ontologies for specific business contexts, purporting the standardization in open data metadata and formats. This offers interesting possibilities for future design science research in the information systems field, among others, semantic modeling as well as knowledge graphs for open data use. While our method synthesizes practitioner knowledge from various open data use cases and input from related open data assessment literature, the exhaustivity of addressed data quality dimensions can be questioned. We primarily considered completeness (on metadata, schema, and dataset content levels), uniqueness, and validity to ensure the usability of open datasets. However, dimensions such as timeliness and accuracy are important for effective open data use. Practitioners in our ADR project repeatedly confirmed that it was more challenging to assess later-known dimensions in an enterprise setting as they required specific business rules and additional validation in the predefined use context, consequently limiting the generalizability of our method. Thus, to thoroughly address the data quality aspects, future research could embed advanced assessment techniques with metrics along additional data quality dimensions in our method's dataset content analysis subphase.

6 Research stream 3: Data sourcing for sustainability

6.1 Motivation and background

As noted earlier, companies are sourcing external data for multiple purposes. However, throughout our research activities of Stream 1, sustainability emerged as one of the most compelling areas of external data sourcing. First, internal factors encourage companies to become more sustainable. These factors include their engagement in more responsible approaches to conduct business, internal audits, and commitments on the part of the boards of directors. This trend is reflected in a recent KPMG survey showing that 96% of the world's top 250 companies have committed themselves to report on their sustainability performance (Threlfall et al., 2020). Second, external factors drive enterprises to rethink their sustainability practices. These factors include strengthened regulatory requirements (Butler, 2011), increased consumers' awareness of the ecologic, social, and economic consequences of their consumption (Lu et al., 2018), and the influence of competitors (Yang, 2018). Third, data availability and data access have emerged as the main issues of sustainability reporting (Deloitte, 2021; EDM Council, 2022; Stoll, 2022). Actually, reporting on sustainability goals is challenging as it requires collecting, processing, and interpreting substantial amounts of data, especially on emissions and product composition, which previously have not yet been systematically collected or analyzed. Even when organizations are able to gather the required data, there is often a lack of detail about its provenance, with the result that they have to rely on estimates.

Despite the relevance of high-quality data to reliably report on sustainability initiatives and goals, there is a void of research on data requirements and data sourcing practices in the context of sustainability. Neither Green IS (Pan et al., 2022; Seidel et al., 2017; Watson et al., 2010) nor environmental management information systems (EMIS), which are supposed to play a significant role in “structured and goal-oriented data gathering, administration, integration, and processing” of environmental information (Stindt et al., 2014), have addressed these topics. Although certain authors highlight data availability and data quality as key issues (Melville et al., 2017; Zampou et al., 2022), they seldom elaborate on data-related topics and only give minimal attention to data accessibility for sustainable development (Machado Ribeiro et al., 2022).

6.2 Research objectives, methodology, and contributions

Building on institutional theory, our research explores the data sourcing practices that companies develop in response to institutional pressures in the sustainability context. Essay 5 is a first step toward the development of a data perspective on sustainability and specifically focuses on data sourcing. We employ institutional theory as a theoretical lens, as it has been widely used in management and sustainability literature (Butler, 2011; Glover et al., 2014; Wang et al., 2015) to study the management practices that enterprises have developed to address regulative, normative, and cultural-cognitive pressures in their environment. We therefore state the following research question: *“How do companies develop data sourcing practices in response to institutional pressures in the sustainability context?”*

In view of our research goals, we leveraged qualitative research methods which are well suited to grasp the richness of specific situations in naturalistic settings (Benbasat et al., 1987; Van de Ven & Poole, 2005). We immersed ourselves in the data sourcing practices of five companies, thereby contributing to in-depth case studies. According to Benbasat et al. (1987), case studies are well suited to capture practitioners' knowledge and develop theories based thereon. Multiple case studies improve external validity while supporting analytical generalization (Yin, 2009). The use of replication logic for study purposes, namely the process of selecting multiple cases that are similar in some important way, allows us to compare and draw conclusions from them (Yin, 2009). For further investigation, we selected five of the 12 companies from our multi-year research. Although all 12 represent large, product-oriented, multinational companies from highly institutionalized industries that currently focus on sustainability goals and commitments, they had reached different levels of maturity in their data sourcing practices and ongoing sustainability initiatives. Using purposeful sampling, we selected the five most mature companies (of the 12) for further investigation. This maturity was reflected by the progress made in their sustainability initiatives and the supporting evidence for a systematic approach to sustainability reporting. Additionally, by selecting five companies representing different industries and positions in the value chain, we expected natural variation with regard to sustainability initiatives and related data sourcing practices, and to better determine the influence of environmental pressures. We collected primary data by conducting semi-structured interviews with key informants – respectively representing each of the five companies. Based on the interviews and secondary data, we then developed process maps for each company and complemented them with additional information about the company's sustainability goals and context of the sustainability initiatives. For the within- and cross-case analyses, we used a research framework, which we have developed by employing institutional theory to analyze and

interpret our empirical insights. The within-case analysis provided a detailed understanding of the unique factors and context that influence the prioritization of sustainability initiatives, namely the motivations behind the engagement, documented in the activities of the planning phase within the process maps. After comprehending the dynamics of each case, we analyzed cross-case patterns to gradually build a rich conceptualization, creating types or groups to compare and examine cases for shared configurations (Miles et al., 2014; Yin, 2009). We employed pattern matching to identify recurring themes across the cases, namely in terms of the exerted pressures and types of the initiatives and the data sourcing practices.

The main contributions of our research are two-fold: First, we propose a research framework based on institutional theory, with which we uncover three data sourcing practices (see Table 5) that companies develop in response to institutional pressures in the sustainability context. Second, our empirical findings include insights into key initiatives in the field of environmental sustainability: understanding the ecological footprint and obtaining labels or complying with regulations on both the product and packaging levels.

Through our cross-case analysis, we identify three general data sourcing practices: sense-making, data collection, and data reconciliation. Sense-making involves the time-consuming analysis of sustainability goals, ambitions, and regulations, and their interpretation in terms of data requirements. Data collection involves the analysis of available data needed to realize the sustainability initiatives, a quality assessment, and gap identification, as well as the collection of missing data from internal and external sources. Data reconciliation encapsulates activities that prepare the data for sustainability reporting. For instance, to calculate the KPIs on the use of recycled material in a specific product, internally and externally collected heterogeneous data should be brought together.

Data sourcing practices	Sense-making	Data collection	Data reconciliation
Activities	Analyze and interpret the sustainability goals and identify the relevant data objects for and attributes of sustainability initiatives Decide on the approach to data collection and processing	Analyze available data needed to implement the sustainability initiatives Assess quality and identify gaps Collect missing data from internal and external sources	Harmonize the definitions and map internal with external reference data. Prepare and aggregate the data for further manipulations and calculations
Outcomes	Relevant data objects and attributes for the sustainability initiative	Quality assessment and gaps in existing data; collection of missing data objects and attributes from internal and external sources	Curated database for KPIs and sustainability reporting
Roles and responsibilities	Sustainability officer, compliance officer, business analyst (sustainability report owner)	Data steward, data analyst, business operations	Data scientist, data engineer
Challenges	<ul style="list-style-type: none"> • Difficulties in adapting to an increasing number of regulations and certifications that address the same SDGs • Interpreting and translating the sustainability goals, legal texts, or certification label requirements into concrete data objects and attributes 	<ul style="list-style-type: none"> • Inability to capture the necessary data along a global supply chain • Missing or erroneous data (e.g., material description) which is presumed to be complete in the enterprise systems 	<ul style="list-style-type: none"> • Heterogeneity of data sources (e.g., variability of types and formats between data from internal and external sources) • Lack of definitions and semantics, as well as difficulties encountered when mapping against them (e.g., recycled material)

Table 5. Data sourcing practices for sustainability

6.3 Discussion, limitations, and outlook

Our study advances sustainability and IS literature by adopting a data perspective and laying the foundation for an academic conceptualization of data sourcing in the context of sustainability. From an academic perspective, our study is an example of impact-oriented Green IS research (Gholami et al., 2016) that guides enterprises on their way to become more sustainable, while embedding sustainability in IS and in practice (Seidel et al., 2017). Regarding the exerted pressures, we find that – in the context of sustainability – the anticipation of regulations instead of the actual regulations as such is the driver of change. The normative and cognitive-cultural pressures are sufficiently prominent to induce companies to act and adapt even before new regulations are promulgated, or when existing regulations persist. It is noteworthy to observe the influence of normative and cognitive pressures where regulations are not yet in existence, while – in other fields – regulations like the UK plastic packaging tax seem to direct sourcing practices despite their inherent complexity. The data sourcing practices suggested in this study

provide a basis for reliable and trustworthy reporting, thereby avoiding or mitigating the risks of greenwashing (Szabo & Webster, 2021). From the perspective of practitioners, the identified sustainability initiatives help them to systematically reflect on data requirements related to sustainability and the need to develop systematic data sourcing practices.

Like most research, this study is not without limitations. First, although we discussed the data sourcing challenges and practices in focus groups involving a larger group of companies that also prioritize other initiatives, our findings are limited to the scope of environmental sustainability initiatives. It would be interesting to replicate our study with initiatives in the fields of social and economic sustainability, thereby enlarging its generalization potential. Second, our findings risk being a “snapshot”. Given that many companies are still in the early phases of their sustainability initiatives and that numerous regulations are expected to be rolled out in future, there are opportunities for longitudinal studies that analyze the evolution of institutional pressures and sourcing practices.

Future research could use the findings of this stream to develop a holistic data sourcing theory that integrates enterprise-wide activities for environmental, social, and economic sustainability. We also see opportunities for academic research that explores how established data management principles and concepts complement data sourcing practices. For instance, such research could explore how data governance can be applied to data sourcing practices to ensure that the sourced data is of a high quality, is properly documented, and is aligned with organizational goals and requirements. More specifically, further research could elaborate which roles can be defined (or adapted) and how they should be integrated to ensure that the sourced sustainability data is effectively managed. Another promising research avenue is data lineage for sustainability data to clarify the origin and movement of sourced data across different systems and processes. Furthermore, regarding the industry setting, the intersection of data sourcing and sustainability undeniably provides exciting opportunities for further enquiries into sustainable supply chains, Green IT, and sustainable computing, as well as for the continued examination of EMIS purporting external data integration.

7 Discussion

7.1 Implications

While the common denominator of data sourcing practices unites the contributions of this thesis, the latter nevertheless has undeniable academic and practical implications that are relevant to the whole IS field, as well as its related subfields.

Enabling the transition from the ad-hoc acquisition of external data to well-informed professional sourcing approaches

Although, in essence, the use of external data is not new and has been mentioned in the enterprise context since the late 1990s (Čas & Meier, 1999; Devlin, 1997), its sourcing has mostly been associated with simply “getting the data”. Irrespective, a systematic approach to the sourcing and managing of external data has been lacking until now. Thus, the outcomes of Essay 2 provide fundamental insights into the way enterprises can perform their data sourcing activities along the six nominal phases and, in the process, allocate the required resources to this task. Accessing, preprocessing, and using external data require an informed decision; a decision that not only considers the price of the datasets, but also relies on a pallet of characteristics covering the transactional, relational, and processual perspectives on data sourcing (see Essay 1).

While Essay 2 focuses on a reference process and underlying design principles, it also provides first insights into emerging roles in the context of external data sourcing and managing by – in the process – distinguishing between two role configurations. In the first and most common role model, existing roles incorporate new activities required by the process. By contrast, the second model presumes that new tasks are taken over by the emerging role of the data hunter or the external data expert. New tasks relating to external data sourcing and managing can either be delegated to a new role or be incorporated into existing roles. We therefore emphasize the imperative that enterprises must implement the missing roles and establish optimal organizational setups to professionalize their external data sourcing approaches.

Improving open data’s readiness for use

Among the discussed challenges of external data and specifically open data (see subsection 5.1) a particular challenge prevails, namely data quality. The quality of published open data has repeatedly been mentioned in academia as a key challenges (Janssen et al., 2012; Stróżyńska et al., 2018; Vetrò et al., 2016); a challenge that is not only addressed by assessment approaches on the

consumers' side, but also by open data providers with a variety of standards and recommendations on how to publish open data of high quality (Berners-Lee, 2012; Data.europa.eu, 2021; Open Data Institute, 2023). Although the overall quality of open datasets in Europe has increased over the past five years (European Commission, 2022b) from a score of 62% to 77% across the 27 EU member states (encompassing the adherence to publishing standards (e.g., DCAT-AP) and also traditional data quality dimensions), the enterprise perception of this quality differs. As the challenges that they face are primarily related to the assessment of open datasets' fitness-for-use, practitioners require proper guidance in this regard. Our findings derived from Essay 4 address these challenges, in the process going beyond the existing assessment methods and showcasing how a well-considered, use-case driven approach can facilitate the search for open datasets and their preparation for use. In addition, the insights gleaned from Essay 3's analysis of open datasets for four specific use cases convincingly demonstrate that even for standardized datasets published by official public authorities – containing high-quality metadata characterized by transparency – the data quality may vary between different providers both within and beyond the confines of the EU. Among others, an important point we make is that such assessments can be done prior to the integration of the data, thus reducing efforts and mitigating risks when using open / external data, and ensuring that the widely discussed benefits of open data are secured.

The close association of sustainability and data sourcing

The sustainability setting, actually, provides a remarkably interesting, highly relevant, and dynamic context to the analysis of data sourcing issues and provides the opportunity to replace ad-hoc practices with more systematic data sourcing practices.

The three identified sourcing practices (see subsection 6.2) outline how companies can transcend ad-hoc approaches when fulfilling sustainability requirements. The findings of Essay 5 highlight that data sourcing for sustainability reporting is inherently more complex than traditional reporting. For instance, in traditional reporting, most of the data is generated internally and managed by accounting teams, whereas in the sustainability context, internal information must be complemented from outside the company (e.g., by suppliers). Another characteristic of data sourcing for sustainability is that data must be repurposed (e.g., product or packaging dimensions) or even created on demand (e.g., prescribing the weight of recycled materials in a product). In this regard, the sense-making derived from internal goals or regulations is a time-consuming and challenging step that requires established collaboration between multiple stakeholders for different practices: sustainability and compliance officers for

sense-making; data stewards, data analysts, and business operations to comply with the sustainability goals and acting on them, as well as to ensure data collection; and data scientists and data engineers for data reconciliation. Sense-making is essential as it clarifies data requirements and identifies data that should be sourced along the global supply chain. Thus, more heterogeneous data comes from various (internal and external) sources, which must be integrated with internal systems and adapted to the new data and business requirements. Further exploring the topic and based on the insights gained from the cases, we noted that the four sustainability initiatives from Essay 5 rely on similar data objects and attributes. We therefore decided to consolidate the data requirements in the form of a conceptual data model that supports sense-making, data collection, and data reconciliation practices. This model conceptualizes the data requirements with reference to ten relevant data objects and attributes of the identified sustainability initiatives.

7.2 Limitations

In addition to the discussions of the specific limitations of each research stream (see subsections 4.3, 5.3, 6.3), it is worth noting that while contributing to a largely novel field of IS research – data sourcing – this thesis undeniably has certain shortcomings. The collection of five essays in this thesis provides valuable research results for academia and practice; results that were derived from an exceptional research setting focusing on large multinational companies (see subsection 3.2) operating on a global scale. Nevertheless, we admit a bias toward this specific type of enterprise (global-scale, multinational companies). We recognize that the companies we collaborated with primarily prioritized the optimization of their data sourcing practices. Therefore, they place a greater emphasis on the exploitation of external data resources rather than exploration (Oberländer et al., 2021). As a consequence, the findings and recommendations derived from our research may not consider specific aspects of data-driven innovation, particularly in the context of external data use for business intelligence and analytics (Božič & Dimovski, 2019). In this regard, ambidexterity and its implications for data sourcing practices unveil an enticing area for future research. The use of external data is not exclusive to large multinational companies and, thus, the current sample could potentially be expanded by including medium-sized or small companies. A more diverse sample, being particularly important in qualitative research designs, could lead to a greater variety of perspectives. This would provide a more nuanced understanding of data sourcing. This calls for further research, such as case studies, which would not only increase the richness of evidence in the emerging domain of data sourcing but would also add granularity. For example, such research could

elaborate on how to source and use external data with a focus on specific industries or geographic areas, considering the cultural and contextual factors that may impact on their data sourcing practices. Additionally, case studies could explore how to effectively source specific types of external data (i.e., beyond open data), which may require more specialized skills and knowledge from the enterprises, also considering their specific contextual requirements and capabilities. Since this thesis is among the first works on data sourcing, the suggested improvements could help to increase the richness and depth of evidence in this emerging domain, thus providing valuable insights that can be used by practitioners and policymakers alike.

Arguably, the demonstrations and evaluations for design science artifacts conducted in Essays 1, 2, and 4 provide valuable insights into the efficacy and usability of these artifacts. However, they may not fully capture the range of challenges and issues that may arise in large-scale implementations. To increase the generalizability and applicability of the proposed solutions, future research should focus on large-scale implementations and evaluations of the external data sourcing taxonomy and reference process (see subsection 4.2) in diverse settings (e.g., enterprise size) with different user groups (e.g., data managers or engineers). Regarding the open data method (Essay 4), especially concerning its context-specific elements, our study would benefit from pre-existing reference ontologies for specific business contexts, calling for future research in this area.

7.3 Outlook and future research

While this thesis explores data sourcing and helps to advance it in the enterprise setting, new opportunities and obstacles might arise due to the emergence of new technologies (e.g., artificial intelligence, blockchain, and Internet of Things). In turn, new requirements for data sourcing could reshape and enhance the findings of this thesis project. For instance, regarding blockchain technologies and their potential to establish secure and transparent data sharing across organizations and industries (Swan, 2015), data sourcing would need to consider factors such as the distributed nature of underlying infrastructure, data privacy, as well as possible novelties in respect of data ownership. Data sharing, already being one of the prominent use cases in the context of this thesis (Essay 1), is an interesting avenue for future research, particularly in the context of data spaces that introduce novel forms of sharing and pooling data (Otto & Jarke, 2019). The outcomes of this thesis provide a required toolbox for the investigation of future data sourcing paradigms, especially as seen from the viewpoint of informed sourcing decisions (Essay 1) and their underlying processes (Essay 2).

Another potential area of research on data sourcing is related to data regulations and the underlying data requirements for compliance. Because of increasing data protection regulations (Labadie & Legner, 2019), access to data – being a critical dimension of data sourcing (see subsection 4.2) – may be restricted by or be subject to new rules that must be critically considered. Future research could investigate how organizations can balance their need for data access with the need for compliance and could identify best practices to source data in heavily regulated industries. The development and evaluation of data sourcing frameworks (based on the findings of this thesis) that incorporate legal and regulatory requirements as key considerations could provide guidance to organizations on how to source data in compliance with relevant laws and regulations, while simultaneously taking into account ethical considerations such as data privacy and security. This type of research would complement our research by making parallels with our results and findings on the data sourcing practices that companies have developed in response to the different types of institutional pressures exerted on them, one of which is regulatory in nature.

8 References

- Aaltonen, A., Alaimo, C., & Kallinikos, J. (2021). The Making of Data Commodities: Data Analytics as an Embedded Process. *Journal of Management Information Systems*, 38(2), 401–429.
- Aaltonen, A., & Tempini, N. (2014). Everything Counts in Large Amounts: A Critical Realist Case Study on Data-Based Production. *Journal of Information Technology*, 29(1), 97–110.
- Aaser, M., & McElhaney, D. (2021). *Harnessing the Power of External Data*. McKinsey. <https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/harnessing-the-power-of-external-data>
- Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data Governance: A Conceptual Framework, Structured Review, and Research Agenda. *International Journal of Information Management*, 49, 424–438.
- Ahlemann, F., & Riempp, G. (2008). RefModPM: A Conceptual Reference Model for Project Management Information Systems. *Wirtschaftsinformatik*, 50(2), 88–97.
- Amsterdamer, Y., & Milo, T. (2015). Foundations of Crowd Data Sourcing. *ACM Special Interest Group on Management of Data (SIGMOD) Record*, 43(4), 5–14.
- Arndt, D., & Gersten, W. (2001). External Data Selection for Data Mining in Direct Marketing. *Proceedings of the Sixth International Conference on Information Quality*, 44–61.
- Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). DBpedia: A Nucleus for a Web of Open Data. *The Semantic Web*, 4825, 722–735.
- Bachtiar, A., Suhardi, & Muhamad, W. (2020). Literature Review of Open Government Data. *Proceedings of the 2020 International Conference on Information Technology Systems and Innovation*, 329–334.
- Baecke, P., & Van den Poel, D. (2011). Data Augmentation by Predicting Spending Pleasure Using Commercially Available External Data. *Journal of Intelligent Information Systems*, 36(3), 367–383.
- Baud, N., Frachot, A., & Roncalli, T. (2002). Internal Data, External Data and Consortium Data—How to Mix Them for Measuring Operational Risk. *SSRN Electronic Journal*.
- Becker, J., Algermissen, L., Delfmann, P., & Knackstedt, R. (2002). Referenzmodellierung. *Das Wirtschaftsstudium*, 30(11), 1294–1298.
- Belissent, J. (2019). *The Insights Professional's Guide to External Data Sourcing*. Forrester Research. <https://www.forrester.com/report/The-Insights-Professionals-Guide-To-External-Data-Sourcing/RES139331>
- Benbasat, I., Goldstein, D. K., & Mead, M. (1987). The Case Research Strategy in Studies of Information Systems. *MIS Quarterly*, 11(3), 369–386.
- Berners-Lee, T. (2012). *5-Star Open Data*. <http://5stardata.info/en/>
- Bicevskis, J., Bicevska, Z., Nikiforova, A., & Oditis, I. (2018). Data Quality Evaluation: A Comparative Analysis of Company Registers' Open Data in Four European Countries. *Communication Papers of the 2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 17, 197–204.
- Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data—The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3), 1–22.
- Božič, K., & Dimovski, V. (2019). Business Intelligence and Analytics Use, Innovation Ambidexterity, and Firm Performance: A Dynamic Capabilities Perspective. *The Journal of Strategic Information Systems*, 28(4), 101578.
- Braunschweig, K., Eberius, J., Thiele, M., & Lehner, W. (2012). The State of Open Data: Limits of Current Open Data Platforms. *Proceedings of the 21st International Conference on World Wide Web*.
- Brynjolfsson, E., Hitt, L. M., & Kim, H. H. (2011). *Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?* (SSRN Scholarly Paper ID 1819486). Social Science Research Network.
- Buhl, H. U., Röglinger, M., Moser, F., & Heidemann, J. (2013). Big Data: A Fashionable Topic with(out) Sustainable Relevance for Research and Practice? *Business & Information Systems Engineering*, 5(2), 65–69.
- Butler, T. (2011). Compliance with Institutional Imperatives on Environmental Sustainability: Building Theory on the Role of Green IS. *The Journal of Strategic Information Systems*, 20(1), 6–26.
- Čas, K., & Meier, M. (1999). Integration of Internal and External Data for Marketing Management. In *Evolution and Challenges in System Development* (pp. 489–503). Springer.
- Chandrasekaran, B. (1990). Design Problem Solving: A Task Analysis. *AI Magazine*, 11(4), 59–71.
- Chen, D. Q., Preston, D. S., & Swink, M. (2021). How Big Data Analytics Affects Supply Chain Decision-Making: An Empirical Analysis. *Journal of the Association for Information Systems*, 22(5), 1224–1244.
- Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, 36(4), 1165–1188.
- Chen, Y., Bharadwaj, A., & Goh, K.-Y. (2017). An Empirical Analysis of Intellectual Property Rights Sharing in Software Development Outsourcing. *MIS Quarterly*, 41(1), 131–161.
- Clark, T. D., Zmud, R. W., & Mccray, G. E. (1995). The Outsourcing of Information Services: Transforming the Nature of Business in the Information Industry. *Journal of Information Technology*, 10(4), 221–237.
- Cleven, A., & Wortmann, F. (2010). Uncovering Four Strategies to Approach Master Data Management. *Proceedings of the 43rd Hawaii International Conference on System Sciences*.

- Corsar, D., & Edwards, P. (2017). Challenges of Open Data Quality: More Than Just License, Format, and Customer Support. *Journal of Data and Information Quality*, 9(1), 1–4.
- Data.europa.eu. (2021). *Data.europa.eu Data Quality Guidelines*. Publications Office of the European Union. <https://data.europa.eu/doi/10.2830/79367>
- Debattista, J., Auer, S., & Lange, C. (2016). Luzzu—A Methodology and Framework for Linked Data Quality Assessment. *Journal of Data and Information Quality*, 8(1), 1–32.
- Deloitte. (2021). *The Importance of ESG Data Management: Challenges and Opportunities for the Real Estate Ecosystem*. <https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/financial-services/deloitte-nl-fsi-real-estate-esg-data-management-whitepaper.pdf>
- Devlin, B. (1997). *Data Warehouse: From Architecture to Implementation* (1st ed.). Addison-Wesley Longman Publishing.
- EDM Council. (2022). *ESG Data Management: Asset Owners*. <https://edmcouncil.org/groups-leadership-forums/esg-data-management/>
- Enders, T., Benz, C., Schüritz, R., & Lujan, P. (2020). How to Implement an Open Data Strategy? Analyzing Organizational Change Processes to Enable Value Creation by Revealing Data. *Proceedings of the 28th European Conference on Information Systems*.
- euBusinessGraph. (2019). Ontology for Company Data. *EuBusinessGraph*. <https://www.eubusinessgraph.eu/eubusinessgraph-ontology-for-company-data/>
- European Commission. (2020a). *Open Data Maturity Report 2019*. https://op.europa.eu/publication/manifestation_identifier/PUB_OABE19001ENN
- European Commission. (2020b). *A European Strategy for Data*. <https://digital-strategy.ec.europa.eu/en/policies/strategy-data>
- European Commission. (2020c). *The Economic Impact of Open Data: Opportunities for Value Creation in Europe*. <https://data.europa.eu/en/node/7773>
- European Commission. (2022a). *Data Act | Shaping Europe's Digital Future*. <https://digital-strategy.ec.europa.eu/en/policies/data-act>
- European Commission. (2022b). *The Open Data Maturity Report 2022*. <https://data.europa.eu/en/news-events/news/open-data-maturity-report-2022-out>
- Fettke, P., & Loos, P. (2003). Classification of Reference Models: A Methodology and its Application. *Information Systems and E-Business Management*, 1(1), 35–53.
- Frank, U. (1999). Conceptual Modelling as the Core of the Information Systems Discipline—Perspectives and Epistemological Challenges. *Proceedings of the 5th Americas Conference on Information Systems*, 3.
- Frank, U. (2014). Multilevel Modeling: Toward a New Paradigm of Conceptual Modeling and Information Systems Design. *Business & Information Systems Engineering*, 6(6), 319–337.
- Gholami, R., Watson, R., Molla, A., Hasan, H., & Bjorn-Andersen, N. (2016). Information Systems Solutions for Environmental Sustainability: How Can We Do More? *Journal of the Association for Information Systems*, 17(8), 521–536.
- Glover, J. L., Champion, D., Daniels, K. J., & Dainty, A. J. (2014). An Institutional Theory Perspective on Sustainable Practices Across the Dairy Supply Chain. *International Journal of Production Economics*, 152, 102–111.
- Hopf, K. (2019). *Predictive Analytics for Energy Efficiency and Energy Retailing*. PhD thesis, University of Bamberg.
- Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, 29(4), 258–268.
- Jarvenpaa, S. L., & Markus, M. L. (2020). Data Sourcing and Data Partnerships: Opportunities for IS Sourcing Research. In *Information Systems Outsourcing* (5th ed., pp. 61–79). Springer.
- Jones, M. (2019). What We Talk about When We Talk about (Big) Data. *The Journal of Strategic Information Systems*, 28(1), 3–16.
- Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. SAGE Publications.
- Könning, M., Westner, M., & Strahringer, S. (2019). A Systematic Review of Recent Developments in IT Outsourcing Research. *Information Systems Management*, 36(1), 78–96.
- Kotlarsky, J., Oshri, I., Dibbern, J., & Mani, D. (2018). *MIS Quarterly Research Curation on IS Sourcing*. <https://www.misqresearchcurations.org/blog/2018/6/26/is-sourcing>
- Kruse, F., Schröder, C., & Gómez, J. M. (2021). Data Source Selection Support in the Big Data Integration Process—Towards a Taxonomy. In *Innovation Through Information Systems. WI 2021* (Vol. 48). Springer.
- Kwon, O., Lee, N., & Shin, B. (2014). Data Quality Management, Data Usage Experience and Acquisition Intention of Big Data Analytics. *International Journal of Information Management*, 34(3), 387–394.
- Labadie, C., & Legner, C. (2019). Understanding Data Protection Regulations from a Data Management Perspective: A Capability-Based Approach to EU-GDPR. *Proceedings of the 14th International Conference on Wirtschaftsinformatik*.
- Lacity, M. C., Khan, S. A., & Yan, A. (2016). Review of the Empirical Business Services Sourcing Literature: An Update and Future Directions. *Journal of Information Technology*, 31(3), 269–328.
- Lacity, M. C., Khan, S., Yan, A., & Willcocks, L. P. (2010). A Review of the IT Outsourcing Empirical Literature and Future Research Directions. *Journal of Information Technology*, 25(4), 395–433.

- Legner, C., Pentek, T., & Otto, B. (2020). Accumulating Design Knowledge with Reference Models: Insights from 12 Years' Research into Data Management. *Journal of the Association for Information Systems*, 21(3), 735–770.
- Leonelli, S. (2015). What Counts as Scientific Data? A Relational Framework. *Philosophy of Science*, 82(5), 810–821.
- Lu, Y., Zhao, C., Xu, L., & Shen, L. (2018). Dual Institutional Pressures, Sustainable Supply Chain Practice and Performance Outcome. *Sustainability*, 10(9), 3247.
- Machado Ribeiro, V., Barata, J., & Cunha, P. da. (2022). Sustainable Data Governance: A Systematic Review and a Conceptual Framework. *Proceedings of the 30th International Conference on Information Systems Development*.
- Mathiassen, L. (2002). Collaborative Practice Research. *Information Technology & People*, 15(4), 321–345.
- Melville, N., Saldanha, T. J. V., & Rush, D. E. (2017). Systems Enabling Low-Carbon Operations: The Salience of Accuracy. *Journal of Cleaner Production*, 166, 1074–1083.
- Miles, M. B., Michael, H. A., & Johnny, S. (2014). *Qualitative Data Analysis: A Methods Sourcebook* (3rd ed.). SAGE Publications.
- Nevo, D., & Kotlarsky, J. (2020). Crowdsourcing as a Strategic IS Sourcing Phenomenon: Critical Review and Insights for Future Research. *The Journal of Strategic Information Systems*, 29(4), 101593.
- Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). A Method for Taxonomy Development and Its Application in Information Systems. *European Journal of Information Systems*, 22(3), 336–359.
- Oberländer, A. M., Röglinger, M., & Rosemann, M. (2021). Digital Opportunities for Incumbents – A Resource-centric Perspective. *The Journal of Strategic Information Systems*, 30(3), 101670.
- Open Data Institute. (2023). *Open Data Publishing*. <https://www.theodi.org/project/research-and-development-improving-data-publishing/>
- Open Government Working Group. (2007). *The 8 Principles of Open Government Data*. <https://opengovdata.org/>
- Orlikowski, W. J., & Scott, S. V. (2014). What Happens When Evaluation Goes Online? Exploring Apparatuses of Valuation in the Travel Sector. *Organization Science*, 25(3), 868–891.
- Osagie, E., Waqar, M., Adebayo, S., Stasiewicz, A., Porwol, L., & Ojo, A. (2017). Usability Evaluation of an Open Data Platform. *Proceedings of the 18th Annual International Conference on Digital Government Research*, 495–504.
- Oshri, I., Kotlarsky, J., & Willcocks, L. P. (2015). *The Handbook of Global Outsourcing and Offshoring* (3rd ed.). Palgrave Macmillan.
- Österle, H., & Otto, B. (2010). Consortium Research: A Method for Researcher-Practitioner Collaboration in Design-Oriented IS Research. *Business & Information Systems Engineering*, 2(5), 283–293.
- Otto, B., & Jarke, M. (2019). Designing a Multi-Sided Data Platform: Findings from the International Data Spaces Case. *Electronic Markets*, 29(4), 561–580.
- Pan, S. L., Carter, L., Tim, Y., & Sandeep, M. S. (2022). Digital Sustainability, Climate Change, and Information Systems Solutions: Opportunities for Future Research. *International Journal of Information Management*, 63, 102444.
- Parvinen, P., Pöyry, E., Gustafsson, R., Laitila, M., & Rossi, M. (2020). Advancing Data Monetization and the Creation of Data-based Business Models. *Communications of the Association for Information Systems*, 47(1), 25–49.
- Peppers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, 24(3), 45–77.
- Piccoli, G., & Pigni, F. (2013). Harvesting External Data: The Potential of Digital Data Streams. *MIS Quarterly Executive*, 12, 143–154.
- Pigni, F., Piccoli, G., & Watson, R. (2016). Digital Data Streams: Creating Value from the Real-time Flow of Big Data. *California Management Review*, 58(3), 5–25.
- Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, 1(1), 51–59.
- Rafati, L., & Poels, G. (2015). Towards Model-Based Strategic Sourcing. In *Achieving Success and Innovation in Global Sourcing: Perspectives and Practices* (Vol. 236, pp. 29–51). Springer.
- Roeder, J., Muntermann, J., & Kneib, T. (2020). Towards a Taxonomy of Data Heterogeneity. *Proceedings of the 15th International Conference on Wirtschaftsinformatik*, 293–308.
- Ruijter, E., Grimmelikhuijsen, S., van den Berg, J., & Meijer, A. (2018). Open Data Work: Understanding Open Data Usage from a Practice Lens. *International Review of Administrative Sciences*, 86(1), 3–19.
- Schatsky, D., Camhi, J., & Muraskin, C. (2019). *Data Ecosystems: How Third-Party Information Can Enhance Data Analytics*. Deloitte. https://www2.deloitte.com/content/dam/insights/us/articles/4603_Data-ecosystems/DI_Data-ecosystems.pdf
- Seidel, S., Bharati, P., Fridgen, G., Watson, R., Albizri, A., Boudreau, M.-C., Butler, T., Kruse, L., Guzman, I., Karsten, H., Lee, H., Melville, N., Rush, D., Toland, J., & Watts, S. (2017). The Sustainability Imperative in Information Systems Research. *Communications of the Association for Information Systems*, 40(1), 40–52.
- Sein, M. K., Henfridsson, O., Puro, S., Rossi, M., & Lindgren, R. (2011). Action Design Research. *MIS Quarterly*, 35(1), 37–56.
- Sorescu, A. (2017). Data-Driven Business Model Innovation. *Journal of Product Innovation Management*, 34(5), 691–696.
- Stindt, D., Nuss, C., Bensch, S., Dirr, M., & Tuma, A. (2014). An Environmental Management Information System for Closing Knowledge Gaps in Corporate Sustainable Decision-Making. *Proceedings of the 35th International Conference on Information Systems*.

- Stoll, A. (2022). *ESG in Your Value Chain*. KPMG. <https://home.kpmg/ch/en/home/insights/2022/06/esg-supply-chain.html>
- Strand, M., & Carlsson, S. A. (2008). Provision of External Data for DSS, BI, and DW by Syndicate Data Suppliers. *Proceedings of the 2008 Conference on Collaborative Decision Making: Perspectives and Challenges*, 245–256.
- Strand, M., & Syberfeldt, A. (2020). Using External Data in a BI Solution to Optimise Waste Management. *Journal of Decision Systems*, 29(1), 53–68.
- Strand, M., Wangler, B., & Olsson, M. (2003). Incorporating External Data into Data Warehouses: Characterizing and Categorizing Suppliers and Types of External Data. *Proceedings of the 9th Americas Conference on Information Systems*, 2460–2468.
- Stróżyńska, M., Eiden, G., Abramowicz, W., Filipiak, D., Małyszko, J., & Węcel, K. (2018). A Framework for the Quality-based Selection and Retrieval of Open Data. *Electronic Markets*, 28(2), 219–233.
- Sun, Z., Di, L., Fang, H., Guo, L., Tan, X., Jiang, L., & Chen, Z. (2021). Agro-Geoinformatics Data Sources and Sourcing. In *Agro-Geoinformatics: Theory and Practice* (pp. 41–66). Springer.
- Swan, M. (2015). *Blockchain: Blueprint for a New Economy* (1st ed.). O'Reilly Media.
- Szabo, S., & Webster, J. (2021). Perceived Greenwashing: The Effects of Green Marketing on Environmental and Product Perceptions. *Journal of Business Ethics*, 171(4), 719–739.
- Tallon, P. P. (2013). Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost. *Computer*, 46(6), 32–38.
- Threlfall, R., King, A., Schulman, J., & Bartels, W. (2020). *The Time Has Come*. KPMG Global. <https://home.kpmg/xx/en/home/insights/2020/11/the-time-has-come-survey-of-sustainability-reporting.html>
- Van Alstyne, M. W., Brynjolfsson, E., & Madnick, S. E. (1995). Why Not One Big Database? Principles for Data Ownership. *Decision Support Systems*, 15(4), 267–284.
- Van de Ven, A. H., & Poole, M. S. (2005). Alternative Approaches for Studying Organizational Change. *Organization Studies*, 26(9), 1377–1404.
- Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A Framework for Evaluation in Design Science Research. *European Journal of Information Systems*, 25(1), 77–89.
- Vetrò, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., & Morando, F. (2016). Open Data Quality Measurement Framework: Definition and Application to Open Government Data. *Government Information Quarterly*, 33(2), 325–337.
- Vom Brocke, J. (2007). Design Principles for Reference Modeling: Reusing Information Models by Means of Aggregation, Specialisation, Instantiation, and Analogy. In *Design Principles for Reference Modeling: Reusing Information Models by Means of Aggregation, Specialisation, Instantiation, and Analogy*. IGI Global.
- Vom Brocke, J., & Buddendick, C. (2006). Reusable Conceptual Models—Requirements Based on the Design Science Research Paradigm. *Proceedings of the 1st International Conference on Design Science Research in Information Systems and Technology*, 576–604.
- Wadmann, S., Johansen, S., Lind, A., Birk, H. O., & Hoeyer, K. (2013). Analytical Perspectives on Performance-based Management: An Outline of Theoretical Assumptions in the Existing Literature. *Health Economics, Policy and Law*, 8(4), 511–527.
- Wang, X., Brooks, S., & Sarker, S. (2015). A Review of Green IS Research and Directions for Future Studies. *Communications of the Association for Information Systems*, 37(1), 395–429.
- Watson, R. T., Boudreau, M.-C., & Chen, A. J. (2010). Information Systems and Environmentally Sustainable Development: Energy Informatics and New Directions for the IS Community. *MIS Quarterly*, 34(1), 23–38.
- Welle Donker, F., & Van Loenen, B. (2017). How to Assess the Success of the Open Data Ecosystem? *International Journal of Digital Earth*, 10(3), 284–306.
- Winter, J. S., & Davidson, E. (2019). Big Data Governance of Personal Health Information and Challenges to Contextual Integrity. *The Information Society*, 35(1), 36–51.
- Winter, R., & Schelp, J. (2006). Reference Modeling and Method Construction: A Design Science Perspective. *Proceedings of the 2006 ACM Symposium on Applied Computing*, 1561–1562.
- Wixom, B. H., & Ross, J. W. (2017). How to Monetize Your Data. *MIT Sloan Management Review*, 58(3), 9–13.
- Yang, C.-S. (2018). An Analysis of Institutional Pressures, Green Supply Chain Management, and Green Performance in the Container Shipping Context. *Transportation Research Part D: Transport and Environment*, 61, 246–260.
- Yin, R. K. (2009). *Case Study Research: Design and Methods* (4th ed.). SAGE Publications.
- Zampou, E., Mourtos, I., Pramataris, K., & Seidel, S. (2022). A Design Theory for Energy and Carbon Management Systems in the Supply Chain. *Journal of the Association for Information Systems*, 23(1), 329–372.
- Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality Assessment for Linked Data: A Survey. *Semantic Web*, 7(1), 63–93.
- Zhang, R., Indulska, M., & Sadiq, S. (2019). Discovering Data Quality Problems: The Case of Repurposed Data. *Business & Information Systems Engineering*, 61(5), 575–593.
- Zhang, S., Zhang, C., & Yang, Q. (2003). Data Preparation for Data Mining. *Applied Artificial Intelligence*, 17(5–6), 375–381.

- Zhao, J. L., Fan, S., & Hu, D. (2014). Business Challenges and Research Directions of Management Analytics in the Big Data Era. *Journal of Management Analytics*, 1(3), 169–174.
- Zrenner, J., Hassan, A. P., Otto, B., & Marx Gómez, J. C. (2017). Data Source Taxonomy for Supply Network Structure Visibility. *Proceedings of the Hamburg International Conference of Logistics*, 117–137.
- Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R., & Alibaks, R. S. (2012). Socio-Technical Impediments of Open Data. *Electronic Journal of E-Government*, 10(2), 156–172.
- Zuiderwijk, A., Janssen, M., Poulis, K., & van de Kaa, G. (2015). Open Data for Competitive Advantage: Insights from Open Data Use by Companies. *Proceedings of the 16th Annual International Conference on Digital Government Research*, 79–88.

Essay 1

Responding to the Siren Song of External Data: Taxonomical Approach to Data Sourcing

Pavel Krasikov and Christine Legner

Faculty of Business and Economics (HEC), University of Lausanne, Switzerland

First version presented at the pre-ICIS Special Interest Group on Advances in Sourcing (SIG Sourcing) workshop, 2021

Extended version for submission to the Electronic Markets – The International Journal on Networked Business, 2023

Abstract: *Enterprises recognize data as a strategic resource, yet their traditional focus has been on using the data collected inside their organization. Meanwhile, enterprises acknowledge that combining internal and external data is beneficial for improving analytics, enriching business processes, and reducing internal data curation efforts. Nonetheless, external data remains underexploited in enterprises and ad-hoc practices still drive external data sourcing, which leads to inconsistent sourcing decisions. To advance data sourcing practices and shed light on relevant criteria, we propose a taxonomy that comprises eight dimensions, organized around transactional, relational, and processual perspectives on data sourcing. The taxonomy is built in three iterations based on scarce but relevant literature, as well as on the insights gained from analyzing data sourcing practices at nine companies. Our research findings, being among the first works on data sourcing, provide a foundation for scientific inquiry on external data sourcing and guide decision-making in an enterprise context.*

Keywords: Data sourcing, Taxonomy, Enterprise data, External data, Data management

Table of contents

1	Introduction.....	46
2	Related work.....	48
2.1	Three perspectives on data sourcing	48
2.2	Characteristics of data as the sourcing object	50
3	Methodology.....	53
3.1	Research process	53
3.2	Taxonomy development.....	54
3.3	Evaluation	58
4	External data sourcing taxonomy.....	59
4.1	Transactional perspective	59
4.2	Relational perspective.....	62
4.3	Processual perspective.....	63
5	Illustrative application of the taxonomy	67
5.1	Data sourcing scenario 1: UID register	67
5.2	Data sourcing scenario 2: D-U-N-S Business Partner Lookup.....	68
5.3	Data sourcing scenario 3: CDQ Data Sharing Community	69
6	Discussion	71
7	Conclusion and outlook	73
8	References	75
	Appendix 1.....	78

List of figures

Figure 3. Taxonomy development iterations	54
Figure 4. Classification of external datasets.....	78

List of tables

Table 6. Characteristics of the three data sourcing perspectives, based on Jarvenpaa and Markus (2020).....	50
Table 7. Overview of data taxonomies that assist in understanding relevant data sourcing characteristics.....	51
Table 8. Overview of ending conditions along the iterations, based on Nickerson et al. (2013)	58
Table 9. External data sourcing taxonomy.....	59
Table 10. The data sourcing taxonomy demonstration for the UID-register	68
Table 11. The data sourcing taxonomy demonstration for D-U-N-S Business Partner Lookup ..	69
Table 12. The data sourcing taxonomy demonstration for the CDQ Data Sharing Community	70

1 Introduction

In the age of digital transformation, data has become a strategic resource and the cornerstone of new business models, decision-making, and value creation (Buhl et al., 2013; Chen et al., 2017; Provost & Fawcett, 2013; Wixom & Ross, 2017). While companies have traditionally considered data as an internal resource, they become more aware of the wealth of data they can acquire from outside the organization; a topic that has been highlighted in academic literature since the late 1990s (Čas & Meier, 1999; Devlin, 1997, p. 135). Several studies provide evidence of the benefits achieved by combining internal and external data in various use contexts (Baecke & Van den Poel, 2011; Hopf, 2019; Strand et al., 2003; Strand & Syberfeldt, 2020; Zrenner et al., 2017). External is employed, among others, to enhance advanced analytics (Kwon et al., 2014; Zhao et al., 2014), enrich business processes (Baecke & Van den Poel, 2011; Baud et al., 2002; Strand & Syberfeldt, 2020), decrease internal data curation efforts (Cleven & Wortmann, 2010), and create new services (Aaser & McElhaney, 2021; Schatsky et al., 2019). However, despite the rapidly increasing variety and volume of data that is available from outside the organization, external data remains underexploited in enterprises (Davenport et al., 2021). Most enterprises lack transparency about the abundance of data that is accessible through external sources, including open data that is available for free, as well as data provided by commercial data providers or shared with other companies (Kitchin & McArdle, 2016; Strand & Carlsson, 2008). Even when enterprises use external data, they source it in an ad-hoc manner, without clear responsibilities and guidelines (Krasikov et al., 2022). Such use of external data not only leads to risky and inconsistent sourcing decisions, but also to redundancies and higher overall costs.

In line with Jarvenpaa and Markus (2020), we argue that companies should develop a more professional approach to data sourcing. While some attempts have been made to facilitate the selection of datasets in specific contexts (Kruse et al., 2021), a data sourcing decision extends beyond the mere selection of suitable datasets (as the object of sourcing). Similar to IS and IT sourcing, external data sourcing decisions should consider different dimensions, among others, the technical specifications and costs of data acquisition, as well as the relationships between and dependencies on the parties that provide the data (Jarvenpaa & Markus, 2020). Data sourcing also requires contractual structures, which differ from other products and services purchased by enterprises, and the management of the entire process, from finding and obtaining the data to integrating and using it. To address these challenges, data sourcing decisions require extensive guidance. However, IS research currently lacks an insightful discussion of the factors that inform these data sourcing decisions and is short of frameworks that can help companies

to assess different sourcing options. To address this gap, we pose the following research question:

RQ: Which dimensions and characteristics inform enterprise data sourcing decisions?

The main result of our research is a taxonomy, being an artifact that facilitates an understanding and analysis of complex phenomena and the making of decisions (Kundisch et al., 2022; Morana et al., 2020; Nickerson et al., 2013). Taxonomies are especially compelling in our field of interest due to their practical relevance and their provision of empirical evaluation (Rizk et al., 2018). To iteratively develop our data sourcing taxonomy, we rely on Nickerson et al.'s (2013) guidelines and leverage the scarce albeit relevant literature and insights gained from data sourcing practices in nine companies. Accordingly, we use deductive (conceptual-to-empirical) and inductive (empirical-to-conceptual) approaches to build the taxonomy. The resulting taxonomy covers three perspectives on data sourcing decisions suggested in the literature, namely transactional, relational, and processual perspectives (Jarvenpaa & Markus, 2020), and concretizes them by outlining eight dimensions and the related characteristics.

Our research contributes to nascent IS literature on data sourcing by proposing – according to Gregor (2006) – a Type 1 theory as “theory for analyzing.” The proposed external data sourcing taxonomy helps to structure the sourcing decisions and systematically assess different sourcing options. As such, it provides a foundation for scientific inquiry into data sourcing by assembling fragmented academic literature and incorporating practitioners’ insights. Practitioners can apply this taxonomy to understand the relevant dimensions and characteristics of and better anticipate challenges inherent to different types of external data. It can furthermore be used as a supporting tool to inform practitioners’ sourcing activities and to assist them in adapting their enterprise-wide data sourcing strategies.

The remainder of this paper is structured as follows: First, we summarize relevant research that provides a better understanding of current perspectives on data sourcing and present an overview of existing literature on applicable taxonomies, thereby elaborating on the research gap. Second, we describe our research approach by detailing the conducted iterations in the taxonomy development process. Third, we present our findings and, through an illustrative application of our taxonomy, elaborate on the data sourcing use cases in enterprises. Finally, we summarize our findings, point out limitations, and suggest avenues for future research.

2 Related work

2.1 Three perspectives on data sourcing

Despite the increasing demand for external data, data sourcing has not been discussed extensively in the literature and, instead, is simply regarded as getting the data (Čas & Meier, 1999; Strand et al., 2003) or selecting suitable datasets for a specific context (Kruse et al., 2021). Building on the concept of IS sourcing, which implies contracting or delegating IS- or IT-related work (Kotlarsky et al., 2018), Jarvenpaa and Markus (2020) made a first attempt to conceptualize the phenomenon and subsequently defined data sourcing as “procuring, licensing, and accessing data (e.g., an ongoing service or one-off project) from an internal or external entity (supplier).” Sun et al. (2021), in turn, discussed different forms of data sourcing. First, conventional data sourcing, which refers to obtaining data from a variety of sources and typically involves finding, obtaining/purchasing, assessing, integrating, and using the data. Second, crowd-based data sourcing, which emerges as a “data procurement paradigm that engages Web users to collectively contribute and process information” (Amsterdamer & Milo, 2015). Third, cloud-based data sourcing, which implies that the cloud-stored data is accessed via dedicated platforms such as Amazon Web Services and Microsoft Azure (Sun et al., 2021). Despite these efforts, data sourcing, which is at the center of this study, has not been extensively addressed in IS literature and, therefore, constitutes a void in the enterprise perspective (Krasikov et al., 2022).

According to Jarvenpaa and Markus (2020), most IS scholars see data as a structured and homogeneous good and therefore consider data sourcing as unproblematic. They view the data in terms of databases, data traces, and data records (Chen et al., 2017; Pigni et al., 2016; Tallon, 2013) that exist in different formats and that are merely used as a commodity to produce the final product or service. Accordingly, the first and most common way of viewing data in data sourcing activities is the commodity or transactional perspective, which emphasizes the exchange and transactional aspects of data sourcing. It positions data sourcing as a natural and simple endeavor (Jarvenpaa & Markus, 2020), and regards data as an easily harvestable commodity to create value-added services (Piccoli & Pigni, 2013). As a result, transaction costs are seen as the determining factor when making the sourcing decision, with a focus on the price of the acquired data, penalties for licensing violations, and access restrictions.

The relational perspective, by contrast, emphasizes the inter-organizational context of data sourcing and the role of external relationships. It views external data sourcing as a strategic

activity that requires trust and collaboration between the enterprise and external parties, where data, as a strategic asset, can be leveraged for competitive advantage through the creation of mutually beneficial partnerships. It also acknowledges that data can be sourced through a variety of organizational arrangements, ranging from bilateral relationships with data providers to multilateral relationships where data is shared and exchanged with peers, which require agreeing on technical data specifications and the related conditions.

The processual perspective of data sourcing, as the third perspective, in turn emphasizes the intra-organizational context and starts by assuming that the sourcing decision and organizational arrangements are control based rather than cost based (Jarvenpaa & Markus, 2020). This perspective goes beyond data acquisition or the transaction as such, and instead considers the entire process from finding and obtaining the data to integrating and using the data, as well as the technical and operational aspects of this process (Krasikov et al., 2022). According to the processual perspective, data sourcing is a complex process that involves several distinct steps and activities, including identifying data needs and goals, identifying potential data sources, evaluating and selecting data sources, and acquiring and integrating data to ensure its use. Table 6 provides an overview of the three perspectives on data sourcing and indicates the key arguments and current state of each in IS research.

Perspective	Description	Main arguments	State of IS research and exemplary topics
Transactional	Emphasizes that external data sourcing is based on transaction cost decisions: with increasing costs to source data, the sourcing decision becomes less attractive (Koutroumpis et al., 2017)	Considers data as a resource, which is easily harvestable and offered “as-is”. Transaction costs are crucial when making the sourcing decision: e.g., price of (acquiring) the data, combination efforts, penalties for licensing violations, and access restrictions.	Prevails in the IS literature (Chen et al., 2017), encompassing the contexts of databases, software programs, data traces, data records, and information artifacts (Abraham et al., 2019; Pigni et al., 2016; Tallon, 2013).
Relational	Discusses the inter-organizational context of how companies approach sourcing decisions and organizational arrangements, based on trusted relationships.	The data is assumed to travel across different use contexts, e.g., in inter-organizational data exchanges (Winter & Davidson, 2019). Relationship and ownership are important value-adding factors for repurposed data (Jarvenpaa & Markus, 2020).	Almost non-existent in IS literature, only being discussed in the context of Big Data governance (Winter & Davidson, 2019) and research data communities (Leonelli, 2015).

Perspective	Description	Main arguments	State of IS research and exemplary topics
Processual	Focuses on the intra-organizational context, where data is “temporal, co-dependent, indeterminant, and pervasively editable” (Jarvenpaa & Markus, 2020). This view also suggests that the sourcing decision and the organizational arrangements are control based rather than cost based.	This view denotes “the value of entanglement of data and operations on data that could take place at any point, from the source to the final reuse” (Jarvenpaa & Markus, 2020). Sourcing processes should be guided by distinctive phases: access, preprocessing, and use (Krasikov et al., 2022).	Emerging in IS literature and primarily reflected in the contexts of electronic medical records (Jones, 2019; Wadmann et al., 2013), social media (Orlikowski & Scott, 2014), digital platforms (Aaltonen & Tempini, 2014), and reference processes to source and manage the external data (Krasikov et al., 2022).

Table 6. Characteristics of the three data sourcing perspectives, based on Jarvenpaa and Markus (2020)

2.2 Characteristics of data as the sourcing object

The characteristics of the sourcing object play a key role in data sourcing decisions, and it is crucial to understand these data characteristics. Although data sourcing has not been discussed extensively in prior research, there have been prior attempts that shed light on data characteristics, including data which originates outside the enterprise (external data). Simmhan et al.’s (2005) taxonomy – to deal with the increase in computational data – pays particular attention to data provenance. Building on this work, Hartig (2009) elaborates on data provenance in the Web and proposes a provenance model for linked data and the underlying metadata documentation. The author suggests capturing the provenance information as a means of justifying data quality and ensuring further reuse. Although not using a rigorous taxonomy development approach, Hopf (2019) advances a taxonomy which explicitly addresses external data and specifies nine categories of external data sources, namely socio-demographic data from address traders, environmental data, public statistical data, geographic data, calendar events, official publications, website content, electronic business platforms, and social media data. Additionally, Kitchin and McArdle (2016) propose a taxonomy of Big Data characteristics, which classifies data sources and compares “small” and “big” data in terms of volume, velocity, and variety. While these works point toward the existence of data sources beyond the studied setting (e.g., enterprises or academia), they mainly focus on (external) data characteristics and do not discuss sourcing aspects.

Various efforts have been made to develop taxonomies for data objects which could indirectly inform data sourcing decision, for instance, to select data sources for a supply network structure and Big Data integration (Kruse et al., 2021), to support data architecture management (Otto et al., 2014), and to explore dataset properties (Roeder et al., 2020). Although the goals and target audience vary, the dimensions proposed by these taxonomies are useful to better understand

the relevant characteristics of the sourced data and to map them to the three perspectives on data sourcing (see Table 7).

Otto et al. (2014) propose a morphological box with 13 dimensions to design and update a data architecture, along with the method for its application. According to them, the “data source” dimension is defined simply as internal or external data. Building on this work, Zrenner et al (2017) suggest a data source taxonomy for supply network structure visibility with 14 dimensions, three of which refer to data sources, namely availability, interface, and pricing model. The remaining dimensions relate to the underlying data properties, including the level of data aggregation, update cycles, ownership, structure, and format. The authors highlight that knowledge on data sources conceptualization is limited, thus justifying their research and its practical application in the automotive industry. Largely building on Zrenner et al’s (2017) taxonomy, Kruse et al. (2021) are among the pioneers of data selection support, focusing on Big Data integration. They almost exclusively concentrate on the technical characteristics of data integration for data science use cases and describe data-related characteristics such as structure, updating, preprocessing, volume, and quality, as well as other characteristics such as licensing and pricing.

Source	Taxonomy name	Target audience and goal	Meta-characteristics / dimensions	Data sourcing perspective
Otto et al. (2014)	Taxonomy of the data resource in the network industry	Designed to provide structure to complex data environments, and to support data architecture management in a networked industry.	m-c: n/a d: 13	Transactional
Zrenner et al. (2017)	Data source taxonomy for supply network structure visibility	Helps practitioners (data managers) to make an initial selection of data sources and provides a general understanding of the data sources.	m-c: n/a d: 14	Transactional
Susha et al. (2017)	Taxonomy of forms of data collaboratives	Distinguishes between different forms of data collaboratives based on how data is shared (supply) and used (demand).	m-c: Data sharing and data supply, data demand, and data use d:14	Relational
Roeder et al. (2020)	Taxonomy of data heterogeneity	Supports researchers and practitioners to explore datasets’ properties.	m-c: Properties of data heterogeneity in diverse data sets d:8	Transactional
Kruse et al. (2021)	The data source taxonomy	Simplifies the data source selection process in the context of Big Data; is practitioner oriented.	m-c: Support the data scientist and decision-maker in data source selection d: 16	Transactional

Table 7. Overview of data taxonomies that assist in understanding relevant data sourcing characteristics

Table 7 summarizes the data taxonomies that assist in understanding relevant characteristics for data sourcing. Overall, we note that several characteristics are repeated in these works, especially access to data, the availability of data, and provenance information. However, only the taxonomy of Zrenner et al. (2017) explicitly named and systematically addressed data sources from outside the organization.

With reference to the three data sourcing perspectives, we find that most data taxonomies cover dataset properties and therefore reflect the relevant dimensions as viewed from the transactional standpoint. Susha et al.'s (2017) taxonomy of collaborative forms adopts a slightly different approach and is deemed to be the only taxonomy that covers the relational perspective. This work distinguishes between two meta-characteristics: data supply and data demand. For data supply, the relevant dimensions are the type of data, the content of data, the diversity of the data providers, and the degree of access to data. Data demand, in turn, focuses on more usage-oriented dimensions in the context of data sharing, namely the target user group, user selection, incentives to use data, collaboration among data users, and the purpose of data use. However, as the goal of this taxonomy is “to convey characteristics of data collaboratives related to data supply (the sharing aspect) and demand for data (the use aspect)” (Susha et al., 2017), it does not guide the data sourcing decision-making process but, instead, clarifies relationships given a data partnership.

To conclude, the scarce literature highlights that data sourcing should embrace several perspectives. It cannot only be considered as a transaction, but should be reflected upon through the prism of relational and processual perspectives (Jarvenpaa & Markus, 2020).

3 Methodology

To address the abovementioned research gap, our aim is to develop a taxonomy to inform enterprises' external data sourcing decisions. Taxonomies are essential tools for both research and practice as they enable scholars and practitioners to comprehend and examine complex domains by classifying their constituent concepts (Nickerson et al., 2013). According to McKnight and Chervany (2001), taxonomies are applied to transform ambiguous concepts into clear concepts by describing their nature and the relationships among them. Furthermore, taxonomies are a form of conceptual knowledge that encompasses both descriptive and prescriptive knowledge (Nickerson et al., 2013).

3.1 Research process

Our research process follows Nickerson et al.'s (2013) method of rigorously building a taxonomy that respects three underlying development criteria. First, we identify meta-characteristics that guide the choice and classification of all the dimensions in the taxonomy; second, we specify end conditions to reach theoretical saturation in the taxonomy; third, we decide on the iterative development of the taxonomy, with the inclusion of both deductive (conceptual-to-empirical) and inductive (empirical-to-conceptual) iterations. In view of the small amount of literature on data sourcing, we argue that both approaches are necessary to build a meaningful taxonomy. In line with Nickerson et al.'s (2013) recommendation, we start with the "conceptual-to-empirical" approach in our first and second iterations, followed by the "empirical-to-conceptual" approach in the third iteration. In doing so, the taxonomy combines theoretical knowledge and empirical findings. In view of the severe lack of literature available on (external) data sourcing, we argue that it is essential to leverage from practical evidence. We did this, as part of an industry-research collaboration on external data sourcing and management, by conducting focus groups and interviews with experts from nine firms. Figure 3 depicts the steps of the taxonomy development process, outlining our research design with reference to the said iterations, approaches, and the evolution of the taxonomy's dimensions.

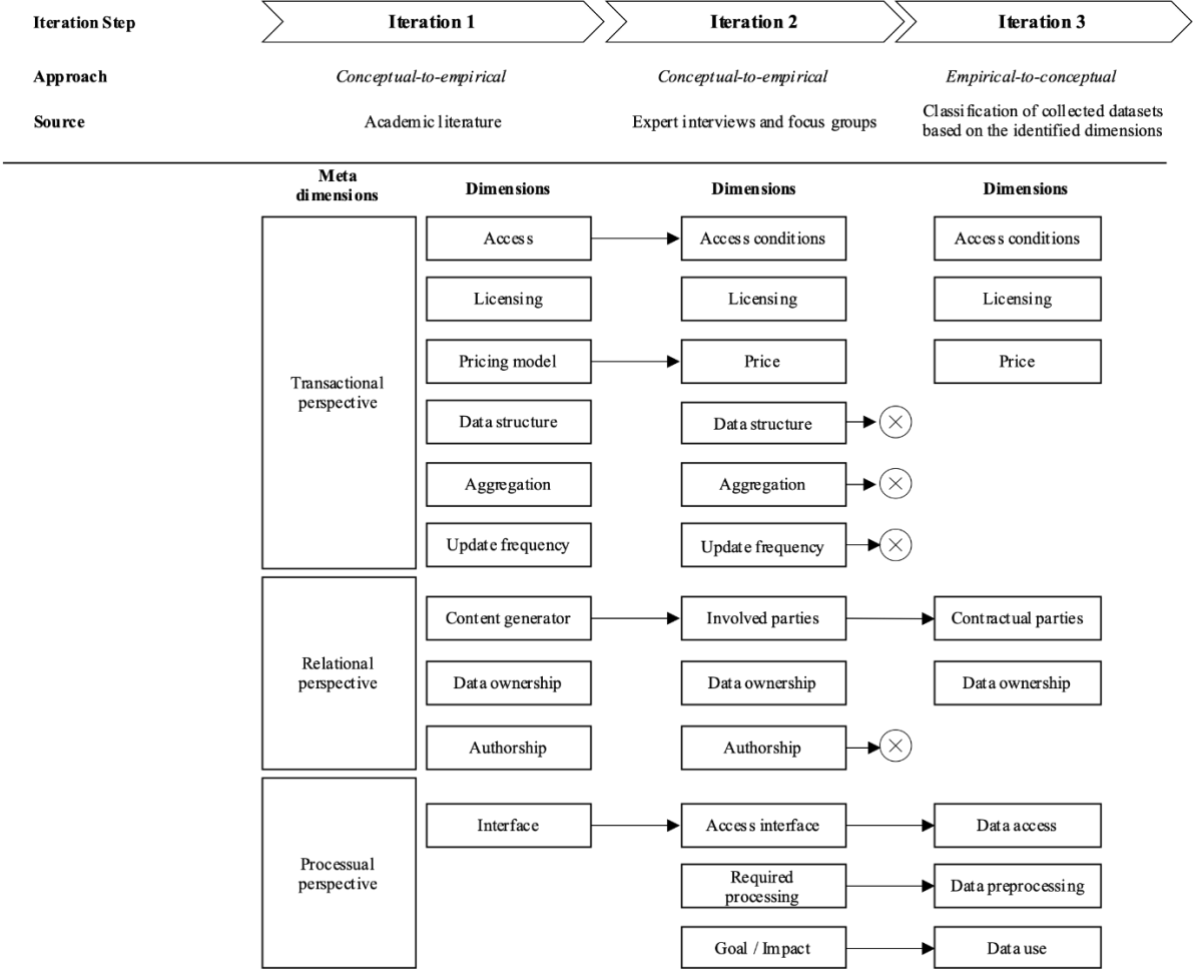


Figure 3. Taxonomy development iterations

It is important to note that several dimensions present in the first iteration were discontinued after receiving feedback following the second iteration, and that new dimensions were added based on input received from practitioners. The modifications of the dimensions are marked either with a cross (depicting elimination) or an arrow (depicting the convergence toward a modified version).

3.2 Taxonomy development

The first step of the taxonomy consists of the definition of the meta-characteristic, i.e., “the most comprehensive characteristic”, a choice that “should be based on the purpose of taxonomy” (Nickerson et al., 2013). Since taxonomies are well-known artifacts in the design science paradigm, meta-characteristics help to embody the identified problem and define “what is relevant for the specific taxonomy design and what is not” (Kundisch et al., 2022). Our choice of a meta-characteristic is determined by the purpose of the taxonomy, which is to inform enterprise data sourcing decision-making. As seen in the section on related work, three

perspectives (i.e., transactional, relational, and processual) are relevant when making sourcing decisions as they enable us to view the data sourcing phenomenon from different angles, especially since “there is no ‘right’ meta-characteristic” (Kundisch et al., 2022).

In the second step, we specify the objective and subjective ending conditions that define when and why the taxonomy is complete and no further iterations are required. Following Nickerson et al.’s (2013) suggestions on when to terminate the design iteration, we build on their extensive list of ending conditions and verify whether they were met by a given version of our taxonomy (see Table 8). For instance, in the case of the objective conditions, it must be ensured that each dimension and characteristic is unique and not repeated, no new dimensions are split, merged, or added during the last iteration, a representative sample of objects is examined, and at least one object is classified under each dimension. As subjective conditions, the taxonomy must be concise, extendible, robust, comprehensive, and explanatory. We indicate whether and which ending condition was met at the end of each iteration.

3.2.1 Iteration 1: conceptual-to-empirical

In accordance with Nickerson et al.’s (2013) taxonomy development method, we relied on three sources to conceptualize the characteristics and dimensions of our taxonomy: scarce literature, experience, and judgement (Kundisch et al., 2022). Since existing taxonomies primarily inform the transactional perspective of data sourcing (see Table 7), we used the taxonomies of Zrenner et al. (2017), Roeder et al. (2020), and Kruse et al. (2021) to gain insights into data source selection and integration and to inform the dimensions of access, price, data structure, aggregation, and update frequency. From the relational perspective, Susha et al.’s (2017) taxonomy of data collaboratives emphasizes the importance of the relationships among the actors in data sharing environments, the subsequent inquiries about data ownership and content generation, and the access interface. From the processual perspective, we did not obtain any input from the literature, except for the interface to access the data which is decisive when it comes to obtaining a desired dataset.

Hence, by examining the scarce literature and building on our expertise in the domain, we deduced the first version of the taxonomy, with dimensions and characteristics in line with the meta-characteristic, to inform the data sourcing decision-making. After verifying the ending conditions (see Table 8) it was evident that although the objective conditions were respected in terms of the uniqueness of attributes and cell duplication (a-c), since no objects were examined for these characteristics, other conditions (d-i) were not met. Regarding the subjective conditions, we were only able to confirm the possibility of adding new dimensions (k), even

though the taxonomy could not yet be considered concise (j), robust (l), or comprehensive (m). Therefore, we continued with a second iteration to revise the current taxonomy.

3.2.2 Iteration 2: conceptual-to-empirical

The objective of our second – conceptual-to-empirical – iteration was to “(re-)examine objects to validate the new characteristics and dimensions” (Kundisch et al., 2022), thereby providing a better understanding of practitioners’ data sourcing decisions. For this purpose, we conducted semi-structured interviews of 30 minutes duration with data experts from nine multinational companies. These companies, based in Germany and Switzerland, were users of external data and represent different industries, among others, pharmaceuticals, manufacturing, transportation, consumer goods, and insurance. Through their involvement in data sourcing activities or initiatives, the key informants were knowledgeable about data that had been sourced and the general data management practices in respective companies. In the first part of the interviews, we gathered information on practitioners’ understanding of external data, namely, the criteria that guide their sourcing decisions. Additionally, we enquired about the current status of the use of external data in the company, as well as about the challenges they encounter when sourcing data and managing the related processes. This enhanced our understanding of the rationale behind their sourcing decisions and enabled us to consider multiple sources for purposes of taxonomy conceptualization (Kundisch et al., 2022). In the second part of the interviews, to “identify a subset of objects to classify” (Nickerson et al., 2013), we collected information on the specific external datasets actually sourced, resulting in a list of 21 datasets and the sources that the companies relied on to obtain them. After removing duplicates, i.e., the same datasets and sources mentioned by several companies, we ended with 16 unique sources. We then returned the list to the interviewees and requested their confirmation of it to ensure the correctness of our documentation. These collected datasets with their sources (ids 1 to 16 in Appendix 1) represent the objects that inform the definition of the taxonomy’s characteristics which, in turn, were grouped in dimensions (Nickerson et al., 2013).

Based on the input obtained through the expert interviews, the examination of the collected objects, and the focus group sessions with participants representing the same companies, we were able to refine the dimensions of the taxonomy. This entailed excluding the data structure, aggregation, and update frequency dimensions from the transactional perspective, since they were less prominent in sourcing decisions and only played a minor role in the actual choice of the data sources. However, the dimensions of access interface, required processing efforts, and the actual goal of sourcing external data, being regarded as crucial for data sourcing decisions by the practitioners, were added to the processual perspective. After verifying the ending

conditions (see Table 8), we realized that the current version of the taxonomy required an additional iteration for convergence. We therefore continued with the third iteration, focusing on empirical evidence.

3.2.3 Iteration 3: empirical-to-conceptual

In our third iteration, we adopted an empirical-to-conceptual approach to further improve the taxonomy and enhance its descriptive purpose (Kundisch et al., 2022), i.e., informing the data sourcing decisions. In line with the general recommendation to conduct at least one iteration of each type (Kundisch et al., 2022; Nickerson et al., 2013), we combined the deductive and inductive approaches, thus examining new objects in this iteration. We expanded the previous iteration's list of collected objects (i.e., external datasets), adhering to a purposeful sampling strategy (Nickerson et al., 2013) based on the availability of information-rich cases of enterprises that provide concrete evidence of sourced datasets. Thus, we collected 23 additional datasets (ids 17 to 39 in Appendix 1) over a span of three years (2019 – 2022) from practitioners participating in executive education courses on data science and management. These datasets represent relevant cases of data sourcing, where companies are actively involved in external data use cases, aimed at analytics, business process improvement, data management, and the creation of new services. Although no new dimensions were added as a result of this iteration, it allowed us to refine the characteristics of contractual parties, data preprocessing, and data use. The complete list eventually comprised 32 sources, nine of which were not considered in the classification exercise due to a lack of evidence of actual use in the enterprise context. With the aim of deploying our revised taxonomy, we then classified a total of 39 sources along the 29 characteristics of the taxonomy. The configuration of similar sets of characteristics, i.e., dimensions, remained intact after the classification exercise, allowing us to verify one of the ending conditions (h, see Table 8). In addition to the previous changes in the taxonomy, we renamed four dimensions (see Figure 3) which were deemed to better correspond to the three perspectives on data sourcing. As a final verification of the ending conditions, we verified mutual exclusivity (f) and collective exhaustivity (g) in addition to the previously confirmed conditions (a-e). All relevant objects from our sample were analyzed (i), allowing us to validate all initially defined objective ending conditions. Most importantly, in terms of subjective conditions and by classifying a newly collected sample in this iteration (ids 17 to 39 in Appendix 1), we confirmed that the taxonomy is comprehensive (m). Finally, we believe that the collected external datasets are sufficiently explained by the taxonomy (n). Nonetheless, we proceeded with an additional evaluation in accordance with the general recommendations for the taxonomy development process (Kundisch et al., 2022; Szopinski et al., 2019).

3.3 Evaluation

The basis of data sources classification was re-examined on three occasions by the authors to assess the ending conditions after each iteration (*ex-ante*). Table 8 provides an overview of the verified ending conditions with regard to our iterative taxonomy development process. To further test the relevance and applicability of the designed taxonomy, we proceeded with an expert evaluation embodied in multiple focus groups, as recommended in Szopinski et al.'s (2019) framework. This *ex-post* quantitative evaluation approach was conducted with a total of 24 experts, representing medium and large Swiss-based companies from various industries, who were not previously involved in the taxonomy development. Three separate focus groups sessions were held between 2019 and 2022 during the abovementioned executive education course. During these sessions, participants were asked to identify, assess, and select external datasets for selected use cases using the dimensions of our taxonomy. This evaluation approach also allowed us to confirm the usability of the data sourcing taxonomy beyond the academic context. Finally, the second version of the taxonomy (Iteration 2) was presented and discussed at the pre-ICIS workshop SIG in Advances in Sourcing 2021 to a group of 15 experts on IS sourcing, enabling us to gather early feedback and to evaluate the academic fit of our findings.

Ending conditions	Iterations		
	1	2	3
<i>Objective conditions</i>			
a) Every dimension is unique, i.e., there are no duplicate dimensions	X	X	X
b) Every characteristic is unique within a dimension	X	X	X
c) No cell duplication, i.e., each cell is unique and not repeated	X	X	X
d) At least one object is classified under every characteristic of every dimension		X	X
e) No split or merger of dimensions, characteristics, or objects occurred		X	X
f) Mutual exclusivity of characteristics, i.e., all characteristics are unique within a dimension			X
g) Collective exhaustivity: for all objects, a characteristic can be assigned to each dimension			X
h) No new dimensions or characteristics were added during the previous iteration			X
i) All relevant objects from the sample were analyzed			X
<i>Subjective conditions</i>			
j) Concise: limited number of characteristics and dimensions		X	X
k) Extendable: possibility to add new dimensions and characteristics	X	X	X
l) Robust: relevance and diversification of identified dimensions and characteristics		X	X
m) Comprehensive: can a random sample of objects within the domain of interest be classified?			X
n) Explanatory: the object is sufficiently explained by dimensions and characteristics			X

Table 8. Overview of ending conditions along the iterations, based on Nickerson et al. (2013)

4 External data sourcing taxonomy

To answer our research question – *Which dimensions and characteristics inform enterprise data sourcing decisions?* – we propose a taxonomy which can be used to evaluate different sourcing options, thus allowing enterprises to make informed decisions on which datasets to source for their intended use purposes. The taxonomy has been iteratively built and its final version comprises eight dimensions distributed along the three dominant perspectives on data sourcing in the IS literature (see Table 9). To provide a useful and meaningful explanation of our taxonomy, we delve deeper into its dimensions and characteristics in the next subsections.

	Dimensions	Characteristics			
Transactional perspective	Access conditions	Open access	Controlled access		Restricted access
	Licensing	License-free	Open-source license		Proprietary license
	Price	Free	One-time payment	Subscription-based	Variable costs
Relational perspective	Contractual parties	Data provider	Data broker		Data intermediary
	Data ownership	Public	Private	Shared	Crowdsourced
Processual perspective	Data access	Web data portal	APIs / Web services	Messages / EDI	File copies
	Data preprocessing	Cleaning	Transformation	Reduction	Integration
	Data use	Analytics	Business process improvement	Data management	Creation of new services

Table 9. External data sourcing taxonomy

4.1 Transactional perspective

The transactional perspective on external data sourcing regards data as a good that can be bought and sold in the market. According to this perspective, sourcing decisions depend on economic aspects, specifically the underlying costs of obtaining the data. These costs are either directly related to the pricing models of external data, or to hidden/indirect costs, particularly when accessing or licensing the external data for further use in the enterprise context. These dimensions prevail in existing data taxonomies concerning external data (see Table 7), three of which have been confirmed by the practitioners as being most relevant in their sourcing decisions.

4.1.1 Access conditions

The access dimension characterizes the conditions of how external data is accessed and which restrictions may apply. For enterprises, well-defined access conditions to the data are crucial to understand whether a dataset can be obtained and accessed for an intended use or not. Zrenner et al. (2017) claim that external data availability should be open or closed. However, these two states are not exhaustive, resulting in the decision to elaborate on the “closed” characteristic as “restricted access” and “controlled access”. While “open” can be interpreted in a various ways, The Open Knowledge Foundation (2005) defines *open access* as being free of any technological restrictions, allowing the seamless reuse of data. Examples of such open datasets were identified in our sample, e.g., the United Nations (UN) dangerous goods list or the Swiss national statistics on gross wages per industry sector. As opposed to open access, datasets can have *restricted access*, which means that the user cannot obtain the data as such without making additional requests and receiving the approval of the provider. When available, *controlled access* requires access permission or authorization and refers to situations where users encounter barriers which complicate but do not restrict the access entirely. For instance, well-known online registration (e.g., account creation or a sign-in with a third-party service) imposes certain conditions which a user must fulfil to gain the desired access to the data. For example, Kaggle (an online community platform which offers different datasets for data science) publishes open datasets, which can be downloaded after registration. In the majority of analyzed datasets from our sample, controlled access is typically related to account creation that makes allowance for additional features offered at the source. In addition, access policies, CAPTCHAs, robot.txt files, or other technological restrictions are considered as restrictions to the access (Open Government Working Group, 2007).

4.1.2 Licensing

Licensing plays an important role in the users’ decisions to source data as it explicitly stipulates legal permissions about the actions that can be undertaken by someone who accesses the data, for instance to share it, to combine it with other data, or to monetize its reuse (Davies, 2012). Three different licensing options prevail for external datasets and are presented as characteristics in our taxonomy. *License-free* specifies that the data is not subject to any reuse restrictions, copyrights, patents, or trademarks (Open Government Working Group, 2007). Although this principle holds true for the public domain, relevant concerns arise regarding copyrighted work, privacy and security, privilege restrictions, and other potential legal concerns. In this regard, specific *open-source* licenses (e.g., Creative Commons) describe alternative ways of how data can be shared or remain available for future use, ranging from most permissive to

most restrictive. This type of licensing is quite prominent in datasets which are made available without any access restrictions (e.g., open data) but which prescribe requirements for licensed data, e.g., attribution, non-commercial use, share-alike, and restrictions on derivative works. Finally, in comparison to the previous categories, specific *proprietary licenses* propose even stricter access to the data under non-standardized conditions. For instance, Nielsen's market research data (an example from our sample) is accompanied by a license agreement on its intended use, requiring permissions for selling or redistribution, which also provides for a code of conduct, ethical standards, and guidelines. These agreed upon conditions guide the further use of external data in the company that plans to source and integrate the data.

4.1.3 Price

From the transactional perspective, pricing is considered as an important dimension when it comes to the choice of external data sources in an enterprise (Arndt & Gersten, 2001), and four different pricing models forge the characteristics of our taxonomy. It comes as no surprise that "most companies purchase lists of information on potential (i.e., new) customers from specialized external vendors" (D'Haen et al., 2013). Zrenner et al. (2017) mention *variable costs*, which can be volume driven or time driven. An example from our collected sample is Lusha (a source for B2B contacts and business leads) whose pricing options depend on the amount of contact detail the user wishes to access over a specific period of time. This example hints at another option, namely *subscription-based fees*, which is a common pricing model. In the example of Lusha, the selected volume of needed contacts can be accessed monthly, annually, or with custom options allowing flexibility in terms of volume and duration. Finally, the *one-time payment* option underpins a widespread pricing scenario, where providers ask for an upfront payment to access the data. For instance, the United Nations Standard Products and Services Code (UNSPSC) codeset can be bought and downloaded. By contrast, situated at the core of open data principles (M. Janssen et al., 2012), there is data which is made available *free of charge*. For instance, many free-of-charge datasets provided by the Swiss Federal Statistical Office were deemed useful by the practitioners that we interviewed. The differences between the pricing methods play an important role in sourcing decisions, particularly concerning the added value behind the invested amount. Although our sample companies indicated a slight preference for open data, this cost-driven logic is unable to account for other factors that drive the sourcing decision.

4.2 Relational perspective

For companies that source external data, it is important to understand the nature of the established relationship with external data provider, data brokers, data marketplaces, and other intermediaries, as well as the ownership rights for the object of sourcing. The reason for this is the transfer of data ownership which occurs at the moment of the sourcing transaction, thereby revealing different ownership paradigms (Loshin, 2001) regarding the intended purpose of use. The relational perspective also recognizes that external data sourcing is not a one-time transaction but an ongoing relationship that requires continuous communication and collaboration between the enterprise and the contractual counterpart.

4.2.1 Contractual parties

Considering that data sourcing primarily involves different parties, typically bound by a contractual relationship (Jarvenpaa & Markus, 2020), we identify three different characteristics to distinguish between the contractual configurations. A commonly established practice in conventional data sourcing is acquisition from a *data provider* (Susha et al., 2017). This contractual party is the entity that provides the data to the client, assuming that it has already collected and stored the data (Sun et al., 2021; Wang et al., 2020). In our sample, this option typically represents a situation where the sourced dataset is generated and provided by the same entity, e.g., national statistical offices, international organizations like the UN (e.g., UNSPSC), or companies (e.g., Homegate with data about Swiss real estate). These entities typically package the data and either sell it or offer freemium models with added services. Strand and Carlsson (2008) note that an external data sourcing decision can also rely on established contractual relations with *data brokers* (also named data suppliers) that acquire data from various sources and process it (if necessary), before selling and delivering the data to prospective clients. We found that data brokers, in particular, are prominent sources of market research data that facilitates an understanding of consumer behavior and preferences, market trends, and related insights, with Mintel, eMarketer, and Nielsen being examples mentioned by the practitioners. Finally, *data intermediary* is an emerging term denoting a contractual party that utilizes a large range of activities to establish a data sourcing relationship, and that is also able to establish relationships with multiple parties to pool data, to enable data sharing, and to establish governance structures (H. Janssen & Singh, 2022). The Humanitarian Data Exchange (HDX), which is essentially a data sharing platform where different organizations share their datasets on humanitarian crises and development issues encompassing topics such as health, education, and demographics, is an example of a data intermediary used by practitioners.

4.2.2 Data ownership

The ownership dimension defines the residual right of control over the data (Van Alstyne et al., 1995) and should be viewed separately from contractually involved parties. The need to consider data ownership in a sourcing decision arises from the fact that ownership determines the legal rights and responsibilities of the parties involved in the sourcing process. By understanding the ownership paradigm that governs the data (Loshin, 2001), relationships can be established that align with the intended use of the sourced data (e.g., creating, reading, consuming, purchasing, compiling, or packaging). The characteristics of this dimension are formed around the main data ownership's socio-organizational notions (Loshin, 2001). For instance, "everyone as owner" promotes the idea of the maximization of benefits and, in line with the accessibility principle, public data goods are available to the *public* (Otto et al., 2014). Although open datasets are typically associated with public ownership, this is not necessarily the norm and ownership would depend on the licensing conditions prescribed by the contractual party. By contrast, if data is privately owned, for instance when "an enterprise creates, processes (adds value), and distributes data about its products" (Fadler & Legner, 2020), then the ownership is *private* (Zrenner et al., 2017). In the latter case, the enterprise cannot exert full control and becomes dependent on the data owner. In addition, owned data can be *shared* by a particular group of individuals or organization as a club good, where multiple entities claim the ownership of the data, e.g., having a shared goal or operating as a business consortium (Susha et al., 2017; Zrenner et al., 2017). Typically, our examples of shared data are embedded in communities (e.g., the CDQ data sharing community, MELLODDY) and involve close collaboration between the parties on a chosen topic of interest or in specific governance mechanisms. Furthermore, the concept of *crowdsourced* data refers to data that is collected, updated, and maintained by a community of Web users who work collectively (Amsterdamer & Milo, 2015). Although the members of this community might not have a direct relationship with one another, the community typically represents a large number of members and assumes the responsibility of "the crowd to generate or source data" (Deutch & Milo, 2012). A well-known example of such data (also used by practitioners in our sample) is the Amazon customer reviews, which are collected on the website via dedicated features.

4.3 Processual perspective

The processual perspective on external data sourcing focuses on the intra-organizational context and the importance of considering the entire data sourcing process as a dynamic, iterative, and evolving activity. By adopting this perspective, enterprises view data sourcing as a series of

interconnected activities, including external data access, preprocessing, and use (Krasikov et al., 2022). External data sourcing decisions are not positioned as a one-time transaction but as an ongoing process that requires continuous monitoring and improvement to ensure that the enterprise effectively uses the external data to realize its business goals. Our finding is that practitioners consider factors such as data access, preprocessing, and use when making a data sourcing decision in the context of this perspective.

4.3.1 Data access

This dimension describes how external data can be accessed from a technological point of view, i.e., the interface as the first point of interaction with the data. Schubert and Legner (2011) distinguish between four types of technical integration, which are reflected as characteristics in the taxonomy. Data can be accessed directly via *web data portals* or *frontends*, which is a typical case when accessing open datasets or social media (e.g., the earlier-mentioned Homegate or Amazon customer reviews). Machine-to-machine coupling, such as messages in electronic data interchange (EDI) – even though rarely encountered – was mentioned in our sample with reference to Descartes, a customs compliance solution that facilitates the reception of timely information on trade complaints and embargos. Another type is the well-known direct links on the database level, namely *file copies*, which are also mentioned by Sucha et al (2017). This type of data access is popular among the entire range of sources, irrespective thereof that it is an open data portal (e.g., *opendata.swiss*, referencing all open data from the Swiss authorities) or a data broker (e.g., GlobalData – business intelligence and market research). Additionally, *APIs* and specifically *web services* are commonly seen as a form of connection to data (Hartig, 2009; Susa et al., 2017), as exemplified by several datasets in our sample. While file copies remain among the most popular means of accessing external datasets (at least in our sample), the ultimate choice depends on the use context. This means that companies can sometimes compare and choose between several options (e.g., web services or file copies) and, therefore, they may seek a source with a specific data access characteristic (e.g., stock listings on the Swiss stock market exchange – SIX) that is crucial in their sourcing decisions. For instance, live data streams, which provide timely updates, cannot conveniently be obtained through downloadable files and are mostly offered via APIs, e.g., real-time data on Swiss road traffic.

4.3.2 Data preprocessing

Our empirical insights highlight that efforts to prepare external data for further use and more specifically the preprocessing of data is an important consideration in the data sourcing context. Data preprocessing can involve *data cleaning*, often described as identifying and removing inconsistencies and outliers from the data, thereby improving the overall quality of the data for

further manipulation (Loshin, 2013). The practitioners mentioned that cleaning efforts are specifically required for datasets with dubious provenance (e.g., crowdsourced data from Amazon customer reviews) and must often be accompanied by additional preprocessing techniques. It is equally important to *transform* external data, originating from heterogeneous sources (Roeder et al., 2020), into a suitable format (e.g., normalize, encode, or scale). Concerning this characteristic, enterprises prefer to source aggregated datasets that consolidate data from different sources because they reduce their own transformation efforts, for instance Geneva's real-time data traffic and parking data (Infomobilité) and real-time Swiss data on road traffic. When it comes to voluminous datasets, particularly in the context of Big Data (Kruse et al., 2021; Roeder et al., 2020) data *reduction* is necessary to reduce dimensionality, remove irrelevant features, or aggregate data to produce a required subset. This is especially useful to produce a substantially smaller dataset for faster processing, while still allowing the user to arrive at the same analytical conclusions. An example of this is the use of data from the Society for Worldwide Interbank Financial Telecommunications (SWIFT) to identify fraudulent transactions, which requires only a fraction of all available information. Finally, when it is necessary to *integrate* external data into the companies' internal systems, schema mapping, merging, joining, and appending are required to bring internal and external data together (Loshin, 2013). In our sample, practitioners aimed to integrate the majority of datasets with their operational or analytics systems, although some were used directly on the web data platform, e.g., HDX.

4.3.3 Data use

The context of external data use is crucial when deciding on suitable data sources. Companies deem it necessary to develop clear use cases for external data, allowing them to plan supporting activities and allocate the necessary resources (Krasikov et al., 2022). Data use not only considers the specific dataset, which is required to be sourced, but also the existing architecture and the capabilities to support and manage the external data sourcing throughout its entire lifecycle. Although the topics and concrete application scenarios vary, academic and practitioner literature distinguishes between four core intentions of external data sourcing in the enterprise context. First, external data is known to enhance *analytics*, particularly when it comes to data enrichment, the improvement of predictive models (Strand & Carlsson, 2008; Wang et al., 2020), or the improvement of the demand forecast with the assistance of external market analytics, data from suppliers, and economic data (Kwon et al., 2014; Zhao et al., 2014). In our sample – in the context of analytics – practitioners mentioned market research data (e.g., GlobalData, Mintel), as well as industry-specific datasets (e.g., Machine Learning Ledger Orchestration for

Drug Discovery – MELLODDY) to accelerate drug discovery. Second, *business process improvement* relies on the combination of internal and external data (Baud et al., 2002; Schatsky et al., 2019). For instance, companies are already using geolocation, weather, and traffic data to plan and optimize their supply chain processes; additional information about exceptional events, such as disasters, enable them to avoid further disruptions of the supply chain (Hopf, 2019; Strand & Syberfeldt, 2020). In this regard, an example drawn from our sample refers to the customs clearance process where the integration of the UN dangerous goods list is intended to optimize and automate document filling. Third, in the context of *data management* as outlined by Cleven and Wortmann (2010), external data can be used to enrich internal data and improve data quality. In these instances, external data creates benefits by reducing data maintenance efforts and improving accuracy through the validation of internal data against external reference data (e.g., the validation of company information with Bureau Van Dijk datasets). Fourth, external data sourcing plays an important role in data-driven innovation, specifically when introducing new products and services to match consumers' needs or to enable the *creation of new services* and business models (Aaser & McElhaney, 2021; Brown, 2021; Shlomo, 2022; Sorescu, 2017). An example of such a service is the people flow management use case, elaborated on by a practitioner, which – based on Swiss Traffic Mobility data – intends to adjust the mobility infrastructure by predicting the flow of people.

5 Illustrative application of the taxonomy

To illustrate the enterprise application of the taxonomy, we selected the external sourcing of business partner data (i.e., data on an enterprise's customers and/or suppliers), being one of the most popular use cases among the sampled datasets of the practitioners. Due to its frequent use in a variety of business processes, including marketing, sales, purchasing, or accounting, business partner data is very critical and needs to be correct and current. Here, the externally sourced business partner data supports master data management and helps to maintain the most accurate version of the data in the company's internal systems; the most prominent attributes being the companies' names, addresses and identifiers (for instance, the value-added tax (VAT) number). The external datasets ensure data quality by removing duplicates, reconciling concepts representing the same real-world object, enriching the data with new entries, and assuring the data's completeness and accuracy by adding up-to-date information obtained from authoritative sources.

In our illustrative application, we consider three data sourcing options for business partner data and demonstrate how the suggested taxonomy supports the sourcing decisions. Data sourcing scenarios 2 and 3 are used by four companies, respectively, while scenario 1 exemplifies the use of openly available data for the validation of business partner records by accessing publicly available information on companies. By comparing the different options, in terms of the taxonomy, we can structure the sourcing decisions and provide empirical evidence of the taxonomy's utility.

5.1 Data sourcing scenario 1: UID register

Among the trustworthy sources of business partner data are company registers as the official data sources published by governments as open data. In this regard, the Swiss UID-register (Unique Enterprise Identification Number) was used by one of the interviewed companies. Company data collected by corporate registers (also known as business registers) has a confirmed reuse potential (Varytimou et al., 2015). In this regard, the UID-register acts as a *data provider* that supplies the datasets to users as *open access* via a dedicated *web data portal*. Therefore, the ownership of the data is *public*. Given that it contains information on the registered companies, the UID-register lists their core properties such as names, full addresses, legal form, status, and related VAT information. The data of the UID-register is provided for *free* under the Swiss Open Government Data License, which is based on the Creative Commons Attribution 4.0 International (CC BY 4.0) *licensing* characteristic. Since the data is provided by

the Swiss Federal Statistical Office, as an authoritative source, it is considered trustworthy and does not need further cleaning, but requires *data integration* efforts to ensure its further use in combination with internal data. Furthermore, the obligation of companies to obtain a UID increases the efficiency of the collaboration between the authorities and the companies, e.g., by standardizing the metadata for facilitated access and the integration of the data. Nevertheless, a company using this dataset to validate business partner records inevitably faces certain challenges regarding the data’s scope (it only contains data on Swiss companies), its unknown quality (e.g., in terms of timeliness), and its maintenance by the data provider.

	Dimensions	Characteristics			
Transactional perspective	Access conditions	Open access	Controlled access		Restricted access
	Licensing	License-free	Open-source license		Proprietary license
	Price	Free	One-time payment	Subscription-based	Variable costs
Relational perspective	Contractual parties	Data provider	Data broker		Data intermediary
	Data ownership	Public	Private	Shared	Crowdsourced
Processual perspective	Data access	Web data portal	APIs / Web services	Messages / EDI	File copies
	Data preprocessing	Cleaning	Transformation	Reduction	Integration
	Data use	Analytics	Business process improvement	Data management	Creation of new services

Table 10. The data sourcing taxonomy demonstration for the UID-register

5.2 Data sourcing scenario 2: D-U-N-S Business Partner Lookup

As a second option, companies can source business partner data from data brokers, such as Dun & Bradstreet (D&B). D&B’s company database comprises information about more than 500 million businesses worldwide and is used by their clients to research prospects, assess customer creditworthiness, or evaluate suppliers (Dun & Bradstreet, 2022). To facilitate its use, this data is typically coupled to specific services such as the identification and classification of data by categories, descriptions of intended use, metadata documentation, and integration services. In the context of *data management*, D&B acts as a *data broker*. It collects company information through the registration procedure assigning the Data Universal Numbering System (D-U-N-S) number, processes the data, and maintains own records for business partner master data. To access D&B data, its clients enter and maintain a contractual relationship with D&B, being the

private owner of the data. While the D-U-N-S search interface preview feature is available for larger user groups in a *web data portal*, customer access to the dataset is only provided under *restricted access* via *APIs / webservice*. This access to the dataset is provided in a *subscription-based* model. Furthermore, to use the D&B dataset, enterprises may need to integrate it with their own databases. This *integration* of the business partner master data can occur through an extended offering with *APIs / web services* upon *subscription* to the paid service.

	Dimensions	Characteristics			
Transactional perspective	Access conditions	Open access	Controlled access	Restricted access	
	Licensing	License-free	Open-source license	Proprietary license	
	Price	Free	One-time payment	Subscription-based	Variable costs
Relational perspective	Contractual parties	Data provider	Data broker		Data intermediary
	Data ownership	Public	Private	Shared	Crowdsourced
Processual perspective	Data access	Web data portal	APIs / Web services	Messages / EDI	File copies
	Data preprocessing	Cleaning	Transformation	Reduction	Integration
	Data use	Analytics	Business process improvement	Data management	Creation of new services

Table 11. The data sourcing taxonomy demonstration for D-U-N-S Business Partner Lookup

5.3 Data sourcing scenario 3: CDQ Data Sharing Community

The third option for sourcing business partner data is a data sharing approach. Four companies identified the CDQ (Corporate Data Quality) Data Sharing Community as their main source of business partner data. In this community, they share business partner data with other companies and try to optimize their data management activities (e.g., business partner master data curation and validation) with the intention of improving data quality and reducing maintenance efforts. The data sharing platform developed by CDQ facilitates the sharing of business partner data among different companies while preserving anonymity, guaranteeing high quality data, and promoting trust in the data. The platform integrates various external data sources and more than 2,100 data quality rules to maintain over 200 million records (CDQ AG, 2023).

From a transactional perspective, the data sharing community provides *restricted access* to companies that are community members. Since CDQ acts as a *data intermediary*, this means that a company cannot access data before entering into a contractual relationship to become a

community member. Licensing conditions are *proprietary*, thus they strictly regulate the use of data beyond the confines of the community to maintain the required privacy level for sensitive data. Although the accessibility of this data source is restrictive, it actually points to the complexity of the multilateral data sourcing relationship, which is more complex than the acquisition of data from a data provider or a broker. CDQ, being a data intermediary, provides access to a platform that pools data from more than 60 external sources (e.g., open government data, corporate registers, and the *shared* data of the community) and provides rule-based data maintenance and validation services (CDQ AG, 2023). The service is accessed through dedicated *APIs / web services*, thus entailing the *integration* of the data into the companies' internal systems such as SAP ERP systems.

	Dimensions	Characteristics			
Transactional perspective	Access conditions	Open access	Controlled access		Restricted access
	Licensing	License-free	Open-source license		Proprietary license
	Price	Free	One-time payment	Subscription-based	Variable costs
Relational perspective	Contractual parties	Data provider	Data broker		Data intermediary
	Data ownership	Public	Private	Shared	Crowdsourced
Processual perspective	Data access	Web data portal	APIs / Web services	Messages / EDI	File copies
	Data preprocessing	Cleaning	Transformation	Reduction	Integration
	Data use	Analytics	Business process improvement	Data management	Creation of new services

Table 12. The data sourcing taxonomy demonstration for the CDQ Data Sharing Community

6 Discussion

As demonstrated by the three sourcing scenarios for business partner data, a variety of sourcing options may exist for the same category of external data. The taxonomy allows an analysis of their implications, viewed from the vantage point of the three data sourcing perspectives.

The **transactional perspective's** dimensions of our data sourcing taxonomy (access conditions, licensing, and price) are among the well-known decision-related factors when sourcing a given dataset (see Table 7), a finding that is in line with traditional sourcing literature (Lacity et al., 2016). Pricing conditions are an important criterion in the selection of specific datasets. Our sample contained a total of 22 datasets that can be accessed for *free*, 20 of which are made available through *open access* while the remaining two can be accessed through registration (*controlled access*). However, even if the data can be accessed as open data without paying additional fees, as in the case of corporate registers like the Swiss UID, its scope is often limited by country. Therefore, companies seeking to validate their records for business partners beyond the borders of Switzerland would have to find and access other open corporate registers. As an alternative, four companies out of our sample preferred using the D-U-N-S business partner lookup, which comes with *subscription-based* costs, but offers a variety of data for the data management use case and has global scope. The third option, CDQ data sharing services, provide subscription-based access to shared data on a global scale, which is *restricted* only to the members of the community. These examples clearly show that the transactional dimensions are not sufficient for a well-informed sourcing decision, because they can neither grasp the complexity underpinning the decisions, nor their consequences, as seen from the relational and processual perspectives.

From the **relational perspective**, the three scenarios allow a better understanding of the nature of contractual relationships and data ownership. It is noteworthy that the contractual agreement is not a prerequisite when sourcing open data from a public institution acting as a *data provider* (e.g., the UID scenario). Here, the public institution, being regulated by an applicable license, neither claims ownership rights over the public data, nor charges fees for any related service. In the case of D-U-N-S business partner lookup, acting as a *data broker*, D&B specifies the terms and conditions, the access to, and the use of the data within the system, as well as any additional services or usage benefits, e.g., assured quality, updates, and privacy. CDQ, being a *data intermediary*, provides access to a platform that enables enterprises to address data quality and maintenance efforts in a collaborative manner. Thus, data ownership can still reside with the

original source, while the data sharing service acts as a facilitator of the data exchange. In this case, the contractual agreement should specify the terms and conditions of the data exchange.

The choice of the data source has several implications relating to data ownership for the contractual parties. In a scenario where a company is using open corporate data, it should be aware that the data is publicly available and that it cannot claim exclusive ownership. However, if a company uses a data sharing service that aggregates data from multiple companies, it is essential to enter into a contractual agreement with the service provider that specifies the ownership and usage rights of the data. Such an agreement also includes clauses on data security, data quality, and data privacy.

From the **processual perspective**, the three sourcing options address the same *data management* use case. Interestingly, while serving the same purpose, as outlined in the data use dimension, the accessing and preprocessing processes differ. External data sourcing, although a complex process and largely driven by the use context (Krasikov et al., 2022), goes through three process steps: data access, data preprocessing, and data use. When using open corporate data from the Swiss corporate register to validate business partner records, the processual perspective involves data access through a dedicated *web data platform* offered by a data provider. Although the interface provides access to wide-ranging data fields (e.g., company name, address, and VAT), the company must still preprocess the data to fit their specific data requirements, i.e., it must perform *cleaning* to ensure the desired level of quality. Therefore, the data preprocessing dimension ensures the accuracy, completeness, and timeliness of the data and provides data quality controls to ensure that the sourced data complies with the company's data standards.

For D&B data, preprocessing mainly includes integration since data quality is part of the contractual agreement and not a company responsibility. For the data sharing scenario, it is similar to the paid service since the contractual agreement allocates accountability to ensure that the sourced data is accessible through a dedicated platform and is ready for use from a quality perspective.

To summarize, considering all three perspectives allows companies to make informed decisions and avoid potential legal, financial, and operational risks associated with external data sourcing.

7 Conclusion and outlook

Our study addresses a gap in current research which, to date, has not addressed data sourcing, despite the increasing relevance of external data. Data sourcing, which aims to identify appropriate sources for data provision, shows similarities with strategic sourcing (Jarvenpaa & Markus, 2020; Krasikov et al., 2022), which aims to select suppliers or vendors providing goods or services. Even though there are clear differences in the object of sourcing, i.e., data versus goods or services, the decisions behind strategic sourcing and data sourcing show similarities and extend beyond the mere exchange or transactional perspective.

Our study contends that sourcing decisions require an analysis of all options through the prism of transactional, relational, and processual perspectives. To support this analysis, we propose an external data sourcing taxonomy, which we built based on a synthesis of scarce literature as well as on the insights gained from analyzing data sourcing practices in nine companies. The taxonomy comprises eight dimensions and 29 characteristics, organized along the three perspectives. Thereby, it outlines complementary angles on what should be considered when assessing sourcing options in the context of external data. It builds on and extends prior studies and existing taxonomies, which investigate data characteristics and inform the transactional perspective. Our empirical insights provide a better understanding of how the taxonomy is used in practice, accompanied by a classification of 39 external datasets used by practitioners in real-life scenarios. We find that the same use case can often involve a variety of sourcing options, that differ not only in the transactional dimensions (*access conditions, licensing and price*), but also in the relational and processual perspectives. Our findings challenge the widely held view that data sourcing is unproblematic, and also draw the attention to the increasing number and variety of sourcing options - ranging from open data to data brokers and marketplaces to the emerging data sharing communities or data collaboratives (Abbas et al., 2021; Bergman et al., 2022; Ruijter, 2021).

From an academic perspective, this paper is among the first attempts to conceptualize data sourcing and advance data sourcing practices in enterprises, by outlining the criteria which drive data sourcing decisions. The proposed taxonomy allows its users to position sourcing options for external data more clearly and it may inspire researchers to elaborate on these options in more detail. The taxonomy-based analysis of sourcing options could be especially interesting for researchers working on open data or the emerging data sharing approaches, as it would help them identify the specific characteristics and challenges from the sourcing perspective. It could also provide insights for data brokers into how to design interesting data products (Davenport

& Kudyba, 2016). In addition, our taxonomy helps to define and analyze sourcing strategies and processes concerning external data in enterprises. It thereby provides valuable guidance for developing professional sourcing practices that help to address the growing demand for external data and cope with the variety of sourcing options. Our findings also contribute to an ongoing discussion of the definition of external data types, especially since the existing definitions vary considerably, even for seemingly well-known types such as open data and its dimensions, social media and online sources, and sensor and IoT data (Kitchin, 2014). Along with *access conditions*, *licensing* and *price*, the same dataset can transition from social media or crowdsourced data (e.g., Twitter or Amazon reviews) to an open dataset published under an open-source license on a web-based data platform (e.g., Kaggle). These changes in the state of different external data types provide an interesting topic for future research since they confirm that data – as the object of sourcing – is far from being a homogeneous good.

From a practitioner' perspective, our taxonomy provides guidance to analyze external data when it is explored or used by a firm, as well as to determine the rationale behind the choice of the data source. We notice that data sourcing is predominantly seen as simply “getting the data” and entering it into systems, while the whole external data sourcing process is actually much more complex, especially as it involves multiple stakeholders (Krasikov et al., 2022)

Although our study is novel and advances the topic of data sourcing in IS literature, it has certain limitations. In line with Nickerson et al. (2013), a taxonomy is useful in a best case but never perfect. Although the suggested dimensions and characteristics are grounded in scarce literature and extensive empirical evidence, our taxonomy (as taxonomies in general) can be extended with additional content by including new discoveries on the use context of external data. For instance, building on evidence derived from existing data taxonomies (see Table 7) and also from the illustrative application of our taxonomy (see Table 10, 6 and 7), we notice the presence and relevance of characteristics which correspond closely to the content of the data (e.g., geographical and temporal coverage, granularity, timeliness and completeness, and structure). While these characteristics undeniably impact on the sourcing decision, they inform an operational decision instead of a strategic decision. In this sense, our proposed taxonomy is a first step toward a more comprehensive taxonomy for external data sourcing.

8 References

- Aaltonen, A., & Tempini, N. (2014). Everything Counts in Large Amounts: A Critical Realist Case Study on Data-Based Production. *Journal of Information Technology*, 29(1), 97–110.
- Aaser, M., & McElhane, D. (2021). *Harnessing the Power of External Data*. McKinsey. <https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/harnessing-the-power-of-external-data>
- Abbas, A. E., Agahari, W., van de Ven, M., Zuiderwijk, A., & de Reuver, M. (2021). Business Data Sharing through Data Marketplaces: A Systematic Literature Review. *Journal of Theoretical and Applied Electronic Commerce Research*, 16(7), 3321–3339.
- Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data Governance: A Conceptual Framework, Structured Review, and Research Agenda. *International Journal of Information Management*, 49, 424–438.
- Amsterdamer, Y., & Milo, T. (2015). Foundations of Crowd Data Sourcing. *ACM Special Interest Group on Management of Data (SIGMOD) Record*, 43(4), 5–14.
- Arndt, D., & Gersten, W. (2001). External Data Selection for Data Mining in Direct Marketing. *Proceedings of the Sixth International Conference on Information Quality*, 44–61.
- Baecke, P., & Van den Poel, D. (2011). Data Augmentation by Predicting Spending Pleasure Using Commercially Available External Data. *Journal of Intelligent Information Systems*, 36(3), 367–383.
- Baud, N., Frachot, A., & Roncalli, T. (2002). Internal Data, External Data and Consortium Data—How to Mix Them for Measuring Operational Risk. *SSRN Electronic Journal*, 1–18.
- Bergman, R., Abbas, A. E., Jung, S., Werker, C., & de Reuver, M. (2022). Business Model Archetypes for Data Marketplaces in the Automotive Industry. *Electronic Markets*, 32(2), 747–765.
- Brown, S. (2021). *Why External Data Should be Part of Your Data Strategy*. MIT Sloan. <https://mitsloan.mit.edu/ideas-made-to-matter/why-external-data-should-be-part-your-data-strategy>
- Buhl, H. U., Röglinger, M., Moser, F., & Heidemann, J. (2013). Big Data: A Fashionable Topic with(out) Sustainable Relevance for Research and Practice? *Business & Information Systems Engineering*, 5(2), 65–69.
- Čas, K., & Meier, M. (1999). Integration of Internal and External Data for Marketing Management. In *Evolution and Challenges in System Development* (pp. 489–503). Springer.
- CDQ AG. (2023). *CDQ: Data Quality Solutions & DQaaS*. <https://www.cdq.com/>
- Chen, Y., Bharadwaj, A., & Goh, K.-Y. (2017). An Empirical Analysis of Intellectual Property Rights Sharing in Software Development Outsourcing. *MIS Quarterly*, 41(1), 131–161.
- Cleven, A., & Wortmann, F. (2010). Uncovering Four Strategies to Approach Master Data Management. *Proceedings of the 43rd Hawaii International Conference on System Sciences*.
- Davenport, T. H., Evgeniou, T., & Redman, T. C. (2021). Your Data Supply Chains Are Probably a Mess. Here's How to Fix Them. *Harvard Business Review*. <https://hbr.org/2021/06/data-management-is-a-supply-chain-problem>
- Davenport, T. H., & Kudyba, S. (2016). Designing and Developing Analytics-Based Data Products. *MIT Sloan Management Review*, 58(1), 83–89.
- Davies, T. (2012). Ten Building Blocks of an Open Data Initiative. *Open Data Impact Blog*. <http://www.opendataimpacts.net/2012/08/ten-building-blocks-of-an-open-data-initiative/>
- Deutch, D., & Milo, T. (2012). Mob Data Sourcing. *Proceedings of the 2012 International Conference on Management of Data*, 581–583.
- Devlin, B. (1997). *Data Warehouse: From Architecture to Implementation* (1st ed.). Addison-Wesley Longman Publishing.
- D'Haen, J., Van den Poel, D., & Thorleuchter, D. (2013). Predicting Customer Profitability During Acquisition: Finding the Optimal Combination of Data Source and Data Mining Technique. *Expert Systems with Applications*, 40(6), 2007–2012.
- Dun & Bradstreet. (2022). *We help clients grow & thrive*. <https://www.dnb.com/en-ch/about-us/what-we-do/>
- Fadler, M., & Legner, C. (2020). Who Owns Data in the Enterprise? Rethinking Data Ownership in Times of Big Data and Analytics. *Proceedings of the 28th European Conference on Information Systems*.
- Gregor, S. (2006). The Nature of Theory in Information Systems. *MIS Quarterly*, 30(3), 611–642.
- Hartig, O. (2009). Provenance Information in the Web of Data. In C. Bizer, T. Heath, T. Berners-Lee, & K. Idehen (Eds.), *Proceedings of the 2009 Workshop on Linked Data on the Web* (Vol. 538).
- Hopf, K. (2019). *Predictive Analytics for Energy Efficiency and Energy Retailing*. PhD thesis, University of Bamberg.
- Janssen, H., & Singh, J. (2022). Data Intermediary. *Internet Policy Review*, 11(1).
- Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, 29(4), 258–268.
- Jarvenpaa, S. L., & Markus, M. L. (2020). Data Sourcing and Data Partnerships: Opportunities for IS Sourcing Research. In *Information Systems Outsourcing* (5th ed., pp. 61–79). Springer.
- Jones, M. (2019). What We Talk about When We Talk about (Big) Data. *The Journal of Strategic Information Systems*, 28(1), 3–16.

- Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. SAGE Publications.
- Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, 3(1), 1–10.
- Kotlarsky, J., Oshri, I., Dibbern, J., & Mani, D. (2018). *MIS Quarterly Research Curation on IS Sourcing*. <https://www.misqresearchcurations.org/blog/2018/6/26/is-sourcing>
- Koutroumpis, P., Leiponen, A., & Thomas, Llewellyn D. W. (2017). *The (Unfulfilled) Potential of Data Marketplaces* (ETLA Working Papers No. 53). <http://pub.etla.fi/ETLA-Working-Papers-53.pdf>
- Krasikov, P., Eurich, M., & Legner, C. (2022). Unleashing the Potential of External Data: A DSR-based Approach to Data Sourcing. *Proceedings of the 30th European Conference on Information Systems*.
- Kruse, F., Schröer, C., & Gómez, J. M. (2021). Data Source Selection Support in the Big Data Integration Process—Towards a Taxonomy. In *Innovation Through Information Systems. WI 2021* (Vol. 48). Springer.
- Kundisch, D., Muntermann, J., Oberländer, A. M., Rau, D., Röglinger, M., Schoormann, T., & Szopinski, D. (2022). An Update for Taxonomy Designers. *Business & Information Systems Engineering*, 64(4), 421–439.
- Kwon, O., Lee, N., & Shin, B. (2014). Data Quality Management, Data Usage Experience and Acquisition Intention of Big Data Analytics. *International Journal of Information Management*, 34(3), 387–394.
- Lacity, M. C., Khan, S. A., & Yan, A. (2016). Review of the Empirical Business Services Sourcing Literature: An Update and Future Directions. *Journal of Information Technology*, 31(3), 269–328.
- Leonelli, S. (2015). What Counts as Scientific Data? A Relational Framework. *Philosophy of Science*, 82(5), 810–821.
- Loshin, D. (2001). Chapter 2: Who Owns Information? In *Enterprise Knowledge Management: The Data Quality Approach* (pp. 25–46). Morgan Kaufmann.
- Loshin, D. (2013). *Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph* (1st ed.). Morgan Kaufmann.
- McKnight, D. H., & Chervany, N. L. (2001). What Trust Means in E-Commerce Customer Relationships: An Interdisciplinary Conceptual Typology. *International Journal of Electronic Commerce*, 6(2), 35–59.
- Morana, S., Pfeiffer, J., & Adam, M. T. P. (2020). User Assistance for Intelligent Systems. *Business & Information Systems Engineering*, 62(3), 189–192.
- Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). A Method for Taxonomy Development and Its Application in Information Systems. *European Journal of Information Systems*, 22(3), 336–359.
- Open Government Working Group. (2007). *The 8 Principles of Open Government Data*. <https://opengovdata.org/>
- Open Knowledge Foundation. (2005). *The Open Definition*. <https://opendefinition.org/>
- Orlikowski, W. J., & Scott, S. V. (2014). What Happens When Evaluation Goes Online? Exploring Apparatuses of Valuation in the Travel Sector. *Organization Science*, 25(3), 868–891.
- Otto, B., Abraham, R., & Schlosser, S. (2014). Toward a Taxonomy of the Data Resource in the Networked Industry. *Proceedings of the 7th International Scientific Symposium on Logistics*, 382–421.
- Piccoli, G., & Pigni, F. (2013). Harvesting External Data: The Potential of Digital Data Streams. *MIS Quarterly Executive*, 12, 143–154.
- Pigni, F., Piccoli, G., & Watson, R. (2016). Digital Data Streams: Creating Value from the Real-time Flow of Big Data. *California Management Review*, 58(3), 5–25.
- Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, 1(1), 51–59.
- Rizk, A., Bergvall-Kärebörn, B., & Elragal, A. (2018). Towards a Taxonomy for Data-Driven Digital Services. *Proceedings of the 51st Hawaii International Conference on System Sciences*. 51st Hawaii International Conference on System Sciences.
- Roeder, J., Muntermann, J., & Kneib, T. (2020). Towards a Taxonomy of Data Heterogeneity. *Proceedings of the 15th International Conference on Wirtschaftsinformatik*, 293–308.
- Ruijter, E. (2021). Designing and Implementing Data Collaboratives: A Governance Perspective. *Government Information Quarterly*, 38(4), 101612.
- Schatsky, D., Camhi, J., & Muraskin, C. (2019). *Data Ecosystems: How Third-Party Information Can Enhance Data Analytics*. Deloitte. https://www2.deloitte.com/content/dam/insights/us/articles/4603_Data-ecosystems/DI_Data-ecosystems.pdf
- Schubert, P., & Legner, C. (2011). B2B Integration in Global Supply Chains: An Identification of Technical Integration Scenarios. *The Journal of Strategic Information Systems*, 20(3), 250–267.
- Shlomo, M. (2022). *Why External Data Is The New Lithium*. Forbes. <https://www.forbes.com/sites/forbestechcouncil/2022/01/31/why-external-data-is-the-new-lithium---and-how-to-get-at-it/>
- Simmhan, Y., Plale, B., & Gannon, D. (2005). A Survey of Data Provenance in e-Science. *ACM Special Interest Group on Management of Data (SIGMOD) Record*, 34(3), 31–36.
- Sorescu, A. (2017). Data-Driven Business Model Innovation. *Journal of Product Innovation Management*, 34(5), 691–696.
- Strand, M., & Carlsson, S. A. (2008). Provision of External Data for DSS, BI, and DW by Syndicate Data Suppliers. *Proceedings of the 2008 Conference on Collaborative Decision Making: Perspectives and Challenges*, 245–256.

- Strand, M., & Syberfeldt, A. (2020). Using External Data in a BI Solution to Optimise Waste Management. *Journal of Decision Systems*, 29(1), 53–68.
- Strand, M., Wangler, B., & Olsson, M. (2003). Incorporating External Data into Data Warehouses: Characterizing and Categorizing Suppliers and Types of External Data. *Proceedings of the 9th Americas Conference on Information Systems*, 2460–2468.
- Sun, Z., Di, L., Fang, H., Guo, L., Tan, X., Jiang, L., & Chen, Z. (2021). Agro-Geoinformatics Data Sources and Sourcing. In *Agro-Geoinformatics: Theory and Practice* (pp. 41–66). Springer.
- Susha, I., Janssen, M., & Verhulst, S. (2017). Data Collaboratives as a New Frontier of Cross-Sector Partnerships in the Age of Open Data: Taxonomy Development. *Proceedings of the 50th Hawaii International Conference on System Sciences*.
- Szopinski, D., Schoormann, T., & Kundisch, D. (2019). Because Your Taxonomy Is Worth It: Towards a Framework for Taxonomy Evaluation. *Proceedings of the 27th European Conference on Information Systems*.
- Tallon, P. P. (2013). Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost. *Computer*, 46(6), 32–38.
- Van Alstyne, M. W., Brynjolfsson, E., & Madnick, S. E. (1995). Why Not One Big Database? Principles for Data Ownership. *Decision Support Systems*, 15(4), 267–284.
- Varytimou, A., Loutas, N., & Peristeras, V. (2015). Towards Linked Open Business Registers: The Application of the Registered Organization Vocabulary in Greece. *International Journal on Semantic Web and Information Systems*, 11(2), 66–92.
- Wadmann, S., Johansen, S., Lind, A., Birk, H. O., & Hoeyer, K. (2013). Analytical Perspectives on Performance-based Management: An Outline of Theoretical Assumptions in the Existing Literature. *Health Economics, Policy and Law*, 8(4), 511–527.
- Wang, W. M., Preidel, M., Fachbach, B., & Stark, R. (2020). Towards a Reference Model for Knowledge Driven Data Provision Processes. In *Boosting Collaborative Networks 4.0* (Vol. 598, pp. 123–132). Springer.
- Winter, J. S., & Davidson, E. (2019). Big Data Governance of Personal Health Information and Challenges to Contextual Integrity. *The Information Society*, 35(1), 36–51.
- Wixom, B. H., & Ross, J. W. (2017). How to Monetize Your Data. *MIT Sloan Management Review*, 58(3), 9–13.
- Zhao, J. L., Fan, S., & Hu, D. (2014). Business Challenges and Research Directions of Management Analytics in the Big Data Era. *Journal of Management Analytics*, 1(3), 169–174.
- Zrenner, J., Hassan, A. P., Otto, B., & Marx Gómez, J. C. (2017). Data Source Taxonomy for Supply Network Structure Visibility. *Proceedings of the Hamburg International Conference of Logistics*, 117–137.

Appendix 1

ID	# of companies using the data	Data source name	Access conditions			Licensing		Price			Contractual parties			Data ownership			Data access			Data preprocessing			Data use													
			Open access	Controlled access	Restricted access	License-free	Open-source license	Proprietary license	Free	One-time payment	Subscription-based	Variable costs	Data provider	Data broker	Data intermediary	Public	Private	Shared	Crowdsourced	Web data platform	APIs / Web-services	Messages / EDI	File copies	Cleaning	Transformation	Reduction	Integration	Analytics	Business process improvement	Data management	Creation of new services					
1	1	NOGA code																																		
2	1	UID Enterprise Identification Number																																		
3	1	UN dangerous goods list																																		
4	4	D-U-N-S business partner lookup																																		
5	1	SWIFT																																		
6	1	Bureau Van Dijk																																		
7	1	MELLODDY																																		
8	1	GlobalData																																		
9	1	Mintel																																		
10	1	eMarketer																																		
11	1	Nielsen																																		
12	1	IRI marketplace																																		
13	1	UNSPSC																																		
14	1	Exchange Data International (EDI)																																		
15	4	CDQ Data Sharing Community																																		
16	1	Amazon customer reviews																																		
17	1	CEDEX Old age and survivors' insurance																																		
18	1	Infomobilité - data about parkings																																		
19	1	Swiss Traffic Mobility																																		
20	1	TARES Commodity code for parts																																		
21	1	Descartes - customs and regulatory																																		
22	1	EIA - international energy market																																		
23	1	CDP - French energy market																																		
24	1	US Income and Product Accounts																																		
25	1	Swiss gross wages by sector																																		
26	1	Swiss demographics by age and canton																																		
27	1	Swiss information on competition																																		
28	1	Swiss nursing staff in the healthcare sector																																		
29	1	Swiss real-time data on road traffic																																		
30	1	Fuel Economy (US)																																		
31	1	Swiss enterprise sizes per industry																																		
32	1	Lusha																																		
33	1	SIX - Swiss stock market																																		
34	1	Homegate Zurich																																		
35	1	Swiss monthly salary in different sectors																																		
36	1	Swiss buildings with occupied dwellings																																		
37	1	Swiss patients from a car accident																																		
38	1	HDX - Ukraine data explorer																																		
39	1	Languages spoken at home in Switzerland																																		

Figure 4. Classification of external datasets

Essay 2

Unleashing the Potential of External Data: A DSR-based Approach to Data Sourcing

Pavel Krasikov, Markus Eurich, and Christine Legner

Faculty of Business and Economics (HEC), University of Lausanne, Switzerland

Published in the Proceedings of the 30th European Conference on Information Systems (ECIS),

2022

Abstract: *External data has become an indispensable pillar in state-of-the-art decision-making and value creation in an enterprise context. Despite the increasing motivation to use external data, information systems (IS) research still lacks an adequate data sourcing perspective. This study aims to address this gap by investigating the practical challenges in this emerging field and developing a reference process for sourcing and managing external data. To this end, we adopt a design science research approach leveraging collaboration with practitioners from nine high-profile companies. Our findings contribute to the scarce body of knowledge on data sourcing in IS by proposing explicit prescriptions in the form of a reference process for sourcing and managing external data.*

Keywords: External data, Data sourcing, Reference process, Design science

Table of contents

- 1 Introduction..... 82
- 2 Related work..... 83
 - 2.1 Sourcing in IS..... 83
 - 2.2 Data sourcing..... 85
 - 2.3 Research gap 86
- 3 Research design and process 87
- 4 Reference process for sourcing and managing external data..... 90
 - 4.1 Challenges and requirements 90
 - 4.2 Reference process..... 91
 - 4.3 Roles 96
- 5 Discussion 98
- 6 Conclusion 99
- 7 References 101

List of figures

Figure 5. Research process	87
----------------------------------	----

List of tables

Table 13. Prior research on data, IT, IS, and strategic sourcing	84
Table 14. Types of data sourcing.....	85
Table 15. Meta-requirements for external data sourcing and managing reference process	91
Table 16. Reference process for sourcing and managing external data.....	92
Table 17. Roles involved in external data sourcing and managing	97

1 Introduction

“With our own data we can only look internally. We need to see industry benchmarks, regional trends, what waves we can ride on; we derive competitive advantage by getting data from outside and enhancing our own data” (Chief Data Officer of Flagstar Bank, as cited in Belissent, 2019). As reflected in this quotation, external data has become essential to practitioners’ decision-making and value-creation processes. A common way to define external data involves the notion of data coming from outside the company (Arndt & Gersten, 2001; Hopf, 2019; Strand & Syberfeldt, 2020). An increasing number of studies show that combining internal data with external data makes it possible to “compete on analytics” (Strand & Carlsson, 2008), enrich business processes, decrease internal data curation efforts, and create new services (Baecke & Van den Poel, 2011; Baud et al., 2002; Schatsky et al., 2019; Strand & Syberfeldt, 2020).

Despite the increasing motivation to use external data, most companies are sourcing data ad-hoc and have not yet established professional sourcing practices. Based on a survey of 100 medium-to-large American companies, the external data provider Explorium (2021) reports that 79% of organizations consider external data to be very valuable. At the same time, 77% of them lack an understanding of external data sourcing processes. Davenport et al. (2021) observe that external data is largely unmanaged within enterprises; they propose that sourcing high-quality data “builds on the process and supplier management techniques used by manufacturers of physical products.” According to Jarvenpaa and Markus (2020), organizations source data for various purposes, but information systems (IS) research – and data management, in particular – lacks a focus on the data sourcing perspective. To address this gap in research, we ask the following two research questions:

What is the current status and challenges in sourcing external data in enterprises?

How should enterprises source and manage external data?

To account for the practical relevance of the topic (March & Smith, 1995), we adopt a design science research (DSR) approach to construct an artifact that “says how to do something,” in line with Gregor (2006)’s type-V theory. Our study is embedded in a multiyear research program in the field of data management, which gives us privileged access to data experts from more than 20 multinational companies. Following the methodological steps suggested by Peffers et al. (2007), we develop a reference process for sourcing and managing external data. Our findings contribute to the scarce literature on data sourcing by proposing explicit prescriptions in the form of a reference process as a generic procedure for evidence-based IS research (Goeken, 2011).

The designed reference process aims to solve the increasingly relevant organizational problems and generalize the process sequence in data sourcing and its elements, such as activities and milestones (Wilmsen et al., 2020).

The remainder of this paper is structured as follows: Section 2 introduces the concept of sourcing in IS research and elaborates on data sourcing. Section 3 outlines our research design and process. Section 4 presents our findings and elaborates on the phases of the reference processes for sourcing and managing external data. In section 5, we summarize and discuss our findings.

2 Related work

The sourcing of external data rarely appears in academic literature, and Jarvenpaa and Markus (2020) have called for research in this domain. In this section, we compare the different types of sourcing (data, IS, IT, and strategic sourcing) with the respective definitions, objects of sourcing, and underlying processes (see Table 13). We conclude that data sourcing approaches resemble existing strategic sourcing processes, but no links have been drawn between the two.

2.1 Sourcing in IS

“Sourcing is the act through which work is contracted or delegated to an external or internal entity that could be physically located anywhere” (Oshri et al., 2015, p. 2). In an enterprise setting, accelerated technological change coupled with the growing importance of supply chain management has helped **strategic sourcing** to evolve from buying to a critical area of strategic management (Rafati & Poels, 2015). Strategic sourcing covers spend analysis, supplier selection and qualification, contract and relationship management, and analytics for the associated decision-making processes.

Sourcing decisions and their success have emerged as fundamental issues in IS sourcing (Watjatrakul, 2005). Kotlarski et al. (2018) define **IS sourcing** as “a broad umbrella term that refers to the contracting or delegating of IS- or IT-related work (e.g., an ongoing service or one-off project) to an internal or external entity (a supplier).” Concerning the mentioned information technology (IT) term, “**IT sourcing** research is a multi-disciplinary research endeavor that examines the organizational impacts of contracting-out IT functions to a third-party provider, from a technology and business perspective” (Sesay & Ramirez, 2016).

Building on the seminal review of the IT outsourcing literature (Lacity et al., 2010), the same authors’ subsequent study of business service sourcing (Lacity et al., 2016) finds consistent evidence that transaction costs (and reduction of costs in general) were determinative for

sourcing decisions. Sourcing is a transaction, but when it comes to terminology, “the words buying, purchasing, procuring, and sourcing are used as synonyms, referring to a transaction where a particular good is transferred between two organizations” (C. O. Schneider et al., 2013). In their review of recent developments in outsourcing in the IT business service, Könning et al. (2019) use “sourcing” and “outsourcing” interchangeably. The authors point to the paradigm shift from cost reduction as a traditional motivator of the sourcing decision, to other stimuli, such as expertise, skill, quality improvement, and focus on core capabilities (Könning et al., 2019). While IS outsourcing decisions have been widely discussed in the literature (Clark et al., 1995; Könning et al., 2019; Lacity et al., 2010; Nevo & Kotlarsky, 2020; Oshri et al., 2015), in the IS infrastructures context (Lyytinen et al., 2017), e.g., hardware and software, the decisions on data sourcing are often seen as a routine taking place in the background (Jarvenpaa & Markus, 2020).

	IS sourcing	IT sourcing	Strategic sourcing	Data sourcing
Definition	“...a broad umbrella term that refers to the contracting or delegating of IS- or IT-related work (e.g., an ongoing service or one-off project) to an internal or external entity (a supplier)” (Kotlarsky et al., 2018)	“...the delegation, through a contractual arrangement, of all or any part of the technical resources, human resources, and the management responsibilities associated with providing IT services to an external vendor” (Clark et al., 1995)	“...a critical area of strategic management that is centered on decision-making regarding an organization’s procurement activities such as spend analysis, capability sourcing, supplier selection and evaluation, contract management and relationship management” (Rafati & Poels, 2015)	“...procuring, licensing, and accessing data (e.g., an ongoing service or one-off project) from an internal or external entity (supplier)” (Jarvenpaa & Markus, 2020)
Sourcing processes	Make the sourcing decision, design contractual structures, and manage the sourcing relationship (Kotlarsky et al., 2018)	IT business services outsourcing processes (Könning et al., 2019)	Identify needs, gather information about relevant factors, evaluate and select suppliers, evaluate best sourcing alternative, contracting (Nordin & Henrik, 2008; Ribas et al., 2021)	Fragmented approaches: Data brokers: Acquire, integrate, assess, sell (Strand & Carlsson, 2008) Clarify data needs, data acquisition, and data application (Wang et al., 2020) Find, assess, decide how to use, understand, obtain/purchase (Sun et al., 2021) Open data: screen, assess, and prepare open data for use (Krasikov et al., 2021)

Table 13. Prior research on data, IT, IS, and strategic sourcing

Jarvenpaa and Markus (2020) argue that, despite the rich body of knowledge on IT and business sourcing services, *data* in IS sourcing research remains unaddressed. Building on the definition by Kotlarski et al. (2018), the authors proclaim **data sourcing** as “procuring, licensing, and accessing data from an internal or external entity” (Jarvenpaa & Markus, 2020).

2.2 Data sourcing

Recent literature (Sun et al., 2021) mentions three major options for data sourcing (see Table 14): conventional sourcing, crowdsourcing, and cloud-based approach.

Types	Definition	Sources
Conventional	Collection of data from a variety of sources, typically involving finding, obtaining/purchasing, assessing, integrating, and using the data.	Strand and Carlsson, 2008; Wang et al., 2020; Krasikov et al. 2021; Sun et al., 2021
Crowd-based	General public (i.e., the crowd) collectively contributes to the generation, aggregation, and processing of the data for its further use.	Deutch and Milo, 2012; Amsterdamer and Milo, 2015; Satish and Yusof, 2017; Lukyanenko et al., 2019; Sun et al., 2021
Cloud-based	Data sourcing from cloud platforms that provide access via dedicated interfaces.	Sun et al., 2021

Table 14. Types of data sourcing

Conventional data sourcing focuses on obtaining data from a variety of sources, which typically involves finding, obtaining/purchasing, assessing, integrating, and using the data. Despite its relevance, it has only played a minor role in the scarce literature on the topic. For instance, Strand and Carlsson (2008) study how data brokers (named syndicate data suppliers) source external data. On a high level, their activities involve acquiring data from various sources, integrating data into internal databases, refining, enriching, and then selling and delivering data to respective clients. This data acquisition perspective adopts a specific example of how external data can be sourced, addressing only the external data made available by specialized data providers. In a recent study, Wang et al. (2020) develop a reference model for knowledge-driven data provision processes in a data engineering environment. The model proposes three key phases: clarification of data needs based on the company's business activities, data acquisition based on the defined requirements and criteria, and data application that would satisfy the knowledge needs. Since the proposed model focuses on supporting data provision in data mining projects, it does not particularly address the acquisition of external datasets from specialized commercial providers. In the specific setting of agro-geoinformatics, Sun et al. (2021) claim that "conventional sourcing depends on human surveyors, is often labor-intensive, and has very tedious administrative processes." The authors highlight the importance of standardization in data sourcing, such as using standard formats and access methods to reach the openly available online resources. Since open data is often positioned among the sourcing candidates for companies (Hopf, 2019; Roeder et al., 2020; Strand & Syberfeldt, 2020; Zuiderwijk et al., 2015), a meta-analysis by Krasikov et al. (2021) reviews the existing studies on open data processes from the publisher and consumer perspective. They define the following core sourcing phases: screen, assess, and prepare open data for use. Overall, we note that although the data

sourcing steps provide a general structure, they focus on specific scenarios of external data use (acquisition of paid sources, knowledge discovery, agro-geoinformatics, or open data) and do not fully address the enterprise perspective on data-sourcing activities.

The concept of crowdsourcing as an “emerging data procurement paradigm that engages Web users to collectively contribute and process information” (Amsterdamer & Milo, 2015) has received more interest in prior IS research. It implies outsourcing tasks to the network of people, such as freelancers, via digital labor marketplaces (Nevo & Kotlarsky, 2020). Thus, it is the responsibility of “the crowd to generate or source data” (Deutch & Milo, 2012). When it comes to companies collecting customer experience data (e.g., Amazon and Netflix), crowdsourcing makes it possible to improve the initial product offering (Satish & Yusof, 2017). Lukyanenko et al. (2019) underscore the value of crowdsourced user-generated content for which companies develop custom information systems.

Similarly, the concept of cloud-based data sourcing stems from the notion of cloud sourcing in IS, which refers to a form of outsourcing of IT resources to the cloud service providers (Lacity et al., 2016; Muhic & Johansson, 2014; S. Schneider & Sunyaev, 2016). Cloud-based data sourcing implies that the data is hosted on the clouds, and access is provided to the user via dedicated interfaces and tools, e.g., Amazon Web Services and Microsoft Azure (Sun et al., 2021).

Concerning external data sourcing, practitioners’ insights should not be neglected. Forrester’s approach to external data sourcing (Belissent, 2019) allocates roles along certain process steps; for instance, data hunters identify and evaluate potential data sources and data architects verify the “fit” of the external data, while procurement draws up the data sourcing contract. Aaser and McElhaney (2021) present McKinsey’s view on how companies should harness the power of external data, where companies are initially advised to establish a dedicated data sourcing team that would take care of finding, accessing, procuring, integrating, reviewing, managing, and using external data.

2.3 Research gap

Compared with the rich body of knowledge on IT (out)sourcing, data sourcing in IS research has not been adequately explored. Some recent developments in the field take a step forward to properly define the concept (Jarvenpaa & Markus, 2020), but a more thorough understanding of the specific challenges associated with data sourcing and the ways to address them is still lacking. Despite the increasing motivation to use external data, data sourcing processes in enterprises and related practices have only occasionally been addressed in the academic literature. In the enterprise setting, only one academic study elaborates on a method addressing

the sourcing processes concentrating on open data (Krasikov et al., 2021). Today, only consulting reports elaborate beyond the nominal sequence of external data sourcing and managing by proposing dedicated roles. Nonetheless, these suggestions remain at the conceptual level and have not been integrated into the academic body of research.

3 Research design and process

To address this gap in research, we adhere to design science research by following the methodology formulated by Peffers et al. (2007) and seek to design actionable guidance for sourcing and managing external data by using a rigorous research process. McCarthy et al. (2020) argue that the successful identification of relevant real-world problems in DSR relies on the engagement of stakeholders (i.e., practitioners) in all phases, starting with the initial problem identification phase. Our multi-year research project debuted in February 2020, when we formed an expert group with practitioners from nine high-profile companies to investigate the challenges related to external data sourcing and management. These practitioners come from different sectors, including pharmaceutical, manufacturing, transportation, consumer goods, and insurance. They were experienced professionals who were already using external data and were involved in the related activities or initiatives in their companies. Figure 5 presents the summary of our research process.

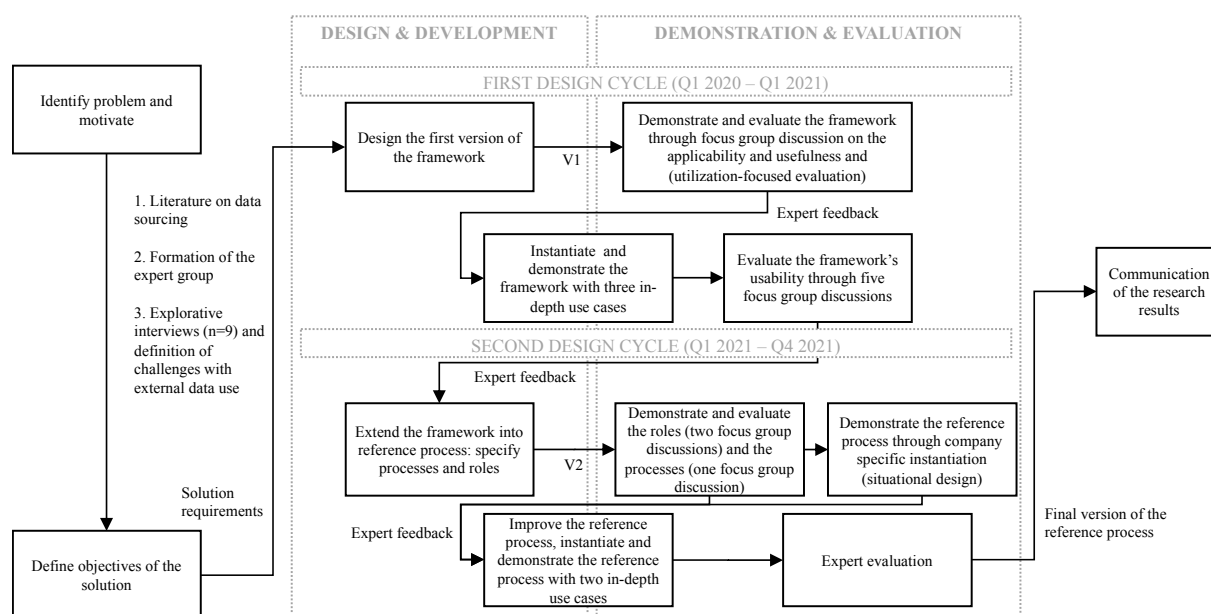


Figure 5. Research process

Following the six steps of the DSR processes model (Peffers et al., 2007) and triggered by an industry need, we chose an objective-centered solution as a research entry point. We started by

identifying the current status of the challenges that practitioners face when using external data. To this end, we conducted nine exploratory semi-structured interviews of 30 minutes with each of the companies mentioned above to understand their initial situation.

In the first design cycle, which lasted from February 2020 to January 2021, we focused on designing an artifact that would help companies address the challenges. This iteration resulted in a framework that depicted the five main phases in external data sourcing and managing. It was presented during a half-day focus group session, where the participants used it to document their own sourcing activities for different types of external data (open, paid, shared, and social media data). The framework was further instantiated for three different use cases, which are highly relevant to the focus group's participants. The first use case relied on using commercial data about companies' affiliations (Dun & Bradstreet, 2018) to map complex organizational relationships within a corporate structure for the company's business partners. The second use case aimed to optimize the logistics process by maintaining dangerous goods classifications and validate it by using openly available sources. The third use case targeted the improvement of corporate sustainability reporting activities with the help of external (shared) data among the different stakeholders (e.g., logistics, procurement, packaging, etc.). These external data use cases were discussed during five separate focus group meetings, which allowed us to consolidate the findings, agree on the applicability and usefulness of the artifact, and initiate a new design phase incorporating the collected feedback.

In the second design cycle (February–November 2021), we included the participants' feedback, refined the main components of the framework, and focused on roles and more granular processes for external data sourcing and management. This cycle marked the transformation from a more generic framework toward a refined artifact. We further demonstrated the applicability of the reference process with two additional use cases of common interest for practitioners. The fourth use case focused on using the public holiday data in sales and marketing analytics to account for the discrepancies between different markets and their geographical granularity. The fifth use case was dedicated to the Web scraping of product reviews from online marketplaces to enhance customer analytics. These scenarios were discussed in three separate focus group meetings. To specify the roles affected by sourcing and managing external data, we analyzed practitioner publications (Aaser & McElhaney, 2021; Belissent, 2019; Schatsky et al., 2019) and conducted four interviews with more experienced companies from the composed expert group. After defining the emerging roles and their corresponding tasks, we held two focus group sessions to understand the additional roles involved in sourcing and managing activities within all the participating companies. Another

session was held to discuss and validate the underlying activities and techniques for each of the main phases of the reference process. Based on these outcomes, one of the authors took part in a company-specific instantiation within one of the expert group companies as a form of naturalistic evaluation (Venable et al., 2012). Based on the framework and the defined activities and roles, the company adapted its own process model for sourcing and managing external data, thereby making it possible to refine the reference process. Specifically, the sourcing phase was split into screening and assessment to further specify the underlying activities of each component. This ex-ante evaluation allowed us to conclude the applicability, utility, and flexibility of our artifact in the enterprise context.

Subsequently, the reference process was further consolidated, and its separate components (phases, activities, roles, and milestones) were discussed and evaluated in four individual sessions with two practitioners from two companies and two external data experts. These two experts have distinctive domain knowledge, while the practitioners have shown interest in applying the findings to the respective companies. The sessions concluded with a questionnaire to evaluate the abovementioned components of the reference process along the typical criteria (Prat et al., 2015) by using a five-point Likert scale. Firstly, we asked the participants to evaluate the relevance of the identified challenges, where the respondents overall agreed with identified challenges. Secondly, the participants were asked to evaluate the relevance of the meta-requirements toward the design of our reference process. The respondents agreed and strongly agreed with the formulated meta-requirements, which are presented below in subsection 4.1. Thirdly, the participants strongly agreed (3/4) that the reference process addresses the real problem and helps manage external data from the sourcing request to the end-of-life, agreed (2/4) and strongly agreed (2/4) that the process ensures that the sourced external data is trusted, compliant, transparent, and of high quality, and agreed (4/4) that it helps clarify roles and responsibilities for sourcing and managing external data. In addition, the reference process was found to be understandable and useful (strongly agree 3/4 and agree 1/4), and applicable to the context of one's company (strongly agree 2/4, agree 1/4, and 1/4 neither agree nor disagree). Fourthly, the participants evaluated the contents of the reference process: They strongly agree (4/4) that the phases of the process are complete, agree (3/4) that roles are appropriate and help clarify the responsibilities for external data, strongly agreed (3/4) that the proposed milestones are appropriate, and shared the opinion that the proposed process variations make the process simpler and more flexible.

4 Reference process for sourcing and managing external data

4.1 Challenges and requirements

Given the benefits of external data use and the increasing interest in developing use cases, companies are making an effort to source external data. The insights from our exploratory interviews showed that practitioners primarily source commercial data sources from the data providers, followed by freely available open data. Nonetheless, our interviewees confirmed that data sourcing processes are, at present, largely unmanaged and subject to multiple challenges (see Table 15). The lack of knowledge about sourcing and managing external data appeared to be the most prominent challenge, coupled with the absence of standards and good practices. Practitioners also emphasized the issue of purchasing external data multiple times within the company, which results in additional efforts and fees. Furthermore, the respondents view the unknown quality of external datasets (e.g., in terms of correctness, completeness, format, consistency) and the lack of trust in external data sources as challenges. In addition, missing knowledge about usage rights, licenses, compliance (e.g., GDPR), and the related legal aspects were troubling. Finally, the lack of transparency about the contents of external data sources raised concerns among the interviewed practitioners. Based on these insights, we formulate the meta-requirements (MR), which are abstract enough to address the class of artifacts but are tied directly to the solution objective:

- MR₁: The artifact should help manage external data from sourcing requests to the end-of-life.
- MR₂: The artifact should clarify roles and responsibilities for external data.
- MR₃: The artifact should ensure that the sourced external data is trusted, compliant, transparent, and high-quality.

Table 15 provides an overview of the challenges with the corresponding meta-requirements. The analysis of the current state and the challenges allowed the expert group to reach a consensus that the outcome of this research should produce a reference process for sourcing and managing external data.

External data sourcing and managing challenges	Meta-requirement
Lack of knowledge about external data sourcing and managing	MR ₁ , MR ₂
No standards for sourcing and managing external data	MR ₁
No overview of which external datasets have been sourced and are used by whom	MR ₁ , MR ₂
Unknown quality of external datasets (e.g., correctness, completeness, format, consistency)	MR ₃
Lack of trust in external data	MR ₃
Missing knowledge about usage rights, licenses, compliance (e.g., GDPR), and legal aspects	MR ₃
Lack of transparency about the contents of external data sources	MR ₃

Table 15. Meta-requirements for external data sourcing and managing reference process

4.2 Reference process

Reference models are typically developed in design and evaluation cycles and are considered important artifacts in IS research (Winter & Schelp, 2006). They were found to be specifically relevant for knowledge accumulation in data management (Legner et al., 2020). In an enterprise setting, a reference process aims to generalize the usual process sequence and its elements, such as activities and milestones (Becker et al., 2007; Wilmsen et al., 2020). To address our specific research objectives, we designed a reference process that provides actionable guidelines supporting companies with external data sourcing and managing.

We iteratively developed and evaluated our reference process for sourcing and managing external data through two major design cycles. Based on the literature regarding data sourcing (see subsection 2.2) and the empirical evidence from the design cycles, we identified six core phases of the process: start, screen, assess, integrate, manage and use, and retire. Each process step contains a clear input, a set of underlying activities, and related roles and techniques. It ends with a defined milestone, allowing a progress review along the reference process. The sequence of the phases is nominal, allowing for the simultaneous execution of activities if the necessary conditions are met. Table 16 provides an overview of the reference process.

	Input	Activities	Techniques	Roles	Output/Milestone
Start	External data request	<ul style="list-style-type: none"> - Define relevant datasets or data needs - Specify the business context - Specify the target system <i>Variants:</i> <ul style="list-style-type: none"> - If use case already exists – skip business requirements - If datasets are known – skip to assessment phase 	<ul style="list-style-type: none"> - Specification of the context requiring new data - Definition of relevant business concepts 	<ul style="list-style-type: none"> - Requestor (data/business) - Data/business analyst - IT specialist 	M1: Documented external data use case with a documentation template
Screen	External data use case	<ul style="list-style-type: none"> - Search for suitable datasets - Identify relevant sources - Locate dataset candidates - Align with data requirements - Check if the data was not already sourced within the organization 	<ul style="list-style-type: none"> - Rely on the metadata provided by the publisher to identify trustable sources - Leverage from dedicated data portals, search engines, and expert knowledge 	<ul style="list-style-type: none"> - Data hunter/data steward 	M2: Candidate datasets identified and documented in a list with names of datasets, publishers, and data sources
Assess	Identified datasets	<ul style="list-style-type: none"> - Define assessment criteria - Verify the budget for the desired datasets - Execute assessment along the defined criteria <i>Variants:</i> <ul style="list-style-type: none"> - If no or not enough datasets fulfilled the criteria – revisit criteria - If no datasets were selected – revisit screening phase 	<ul style="list-style-type: none"> - Three-level assessment: metadata, schema, and content 	<ul style="list-style-type: none"> - Data steward - Procurement - Compliance officer 	M3: Selected datasets, purchase order/contact with data provider
Integrat	Selected datasets	<ul style="list-style-type: none"> - Access external datasets - Identify relevant business concepts - Define the owner of newly identified data 	<ul style="list-style-type: none"> - Rely on knowledge graph mapping approaches 	<ul style="list-style-type: none"> - IT specialist - Data steward 	M4: Integrated external datasets
Manage and use	Integrated external datasets	<ul style="list-style-type: none"> - Establish governance for external data - Use external data - Manage updates - Monitor the use <i>Variants:</i> <ul style="list-style-type: none"> - Continuous use of external data, unless no termination of use case is planned 	<ul style="list-style-type: none"> - Ensure clear guidelines for the use of external data 	<ul style="list-style-type: none"> - Requestor (data / business) - External data expert - Compliance officer 	M5: Decision to terminate the use case
Retire	Termination decision	<ul style="list-style-type: none"> - Decide how external data is treated at the end-of-life <i>Variants:</i> <ul style="list-style-type: none"> - If new data is needed for the use case – back to screening phase - If use case is retired but not the data – back to start phase 	<ul style="list-style-type: none"> - Adhere to a secure data archiving approach within the company 	<ul style="list-style-type: none"> - Data steward - Procurement 	If the dataset is no longer used, it should be retired

Table 16. Reference process for sourcing and managing external data

4.2.1 Start

The first phase is triggered by a request for external data and aims to define and document the motivation for sourcing the new data. As seen in subsection 2.2, clarification of data needs (Wang et al., 2020) makes it possible to initiate the sourcing process and lays the groundwork to define the concrete use case–related requirements and identification of the sourcing goal. The

process originates with the receipt of the external data demand for a potential use case. Since this is an initiation phase, there are several possibilities for how the activities can develop: Variant 1a – the use case does not exist, and the external datasets are unknown; Variant 1b – the use case already exists, but the external datasets are unknown; Variant 1c – the use case already exists, and the datasets are known.

In the first variant, since the use case does not exist, the business requirements for the use of external data should be defined. This includes the trigger and purpose of the use case, as well as responsible functions and teams within the organization. Since the potential external datasets are still unknown, the data requirements must be defined. This comprises potentially relevant sources, their geographical and temporal coverage, as well as their respective levels of granularity. In addition, it is necessary to specify relevant data objects (internal and external business concepts/attributes) that serve as primary data requirements. Furthermore, the system requirements need to be specified, namely the target system for onboarding external data and its required format. These activities enable our reference process and directly address the MR1. Variant 1b becomes relevant when companies have already formulated precise use cases involving external data and can skip the business requirements formulation, starting directly with data requirements. Variant 1c implies that the initiation phase is no longer needed, since the company is assumed to have a clear picture of their business, data, and system requirements already, which allows them to skip directly to the assessment phase (see subsection 4.2.3). The milestone (M1) reached in this phase represents a documented use case for external data, encompassing the requirements mentioned above.

4.2.2 Screen

The screening phase primarily targets the identification of relevant sources and underlying datasets. Finding relevant sources appears to be a challenge beyond our expert group – for instance, 74% of respondents in a survey conducted by Explorium (2021) confirm they are “not sure what to look for.” Searching proves to be even more challenging when it comes to the resources freely available online (Krasikov et al., 2021), as opposed to a more context-specific offering provided by data brokers. Therefore, the input of this phase builds on the previously defined external data use cases, particularly the primary data requirements. As one of the common data sourcing steps (Sun et al., 2021; Wang et al., 2020), screening intends to locate relevant external data sources. These possible sources could include open data portals, traditional search engines or dedicated dataset search engines, data providers and brokers, shared platforms, social media data, and expert knowledge (Krasikov et al., 2021; Roeder et al., 2020; Strand & Carlsson, 2008; Strand & Syberfeldt, 2020). Once candidate sources have been

identified, relevant datasets candidates are located, in line with the primary data requirements of the previous phase. To get a better understanding of the data, dataset samples can be viewed directly at the source level, if available, or requested from the data provider. Moreover, referring to the internal data catalog is crucial to verify whether the data has not already been sourced within the organization. The milestone (M₂) in this phase represents a list with names of the candidate datasets, along with publisher details, to keep record of the source. Furthermore, contracts and service-level agreements (SLA) negotiations must be conducted in line with procurement guidelines.

4.2.3 Assess

This phase aims to assess the different elements of the data based on the predefined selection criteria. The distinction in the assessment criteria relies on whether external data is paid or non-paid. Namely, our insights from the focus group and the evidence from Belissent's (2019) guide show that paid data requires the procurement team's involvement to select the datasets that align with the allocated budget and fulfill the use case requirements. Based on the feedback from the focus group discussion, we adopt a multifaceted assessment approach (Krasikov et al., 2021), since it addresses the core data-related challenges (see subsection 4.1) and proposes three levels of assessment. The first level (metadata) conveys a multitude of information, such as access conditions, licensing information (permissions and prohibitions), publishing details (publisher, publishing date, update cycle), and general content-related information (language, geographical and temporal coverage, number of records and attributes). The second (schema level) is important to verify whether the needed attributes (see subsection 4.2.1) are present in the preselected external datasets. The third level (content assessment) investigates the dataset contents in terms of typical data quality dimensions and their metrics, such as completeness, uniqueness, and validity. We identify three possible variants of how the assessment criteria can be met: Variant 3a – assessment criteria are met, and the necessary number of datasets is selected; Variant 3b – no (or not enough) datasets passed assessment criteria, and the assessment criteria should be revisited; Variant 3c – no datasets passed the assessment criteria, and the screening phase should be revisited.

This phase concludes with the achievement of M₃ with a list of selected datasets that have passed the defined assessment criteria. The process flexibility brought about by these variations is important since the evidence shows the quality of external data may vary. Thus, the selection criteria are subject to revisions on the company's side (Variant 3b) to make the use case feasible. If no datasets pass the assessment criteria, the screening phase must be revisited to identify further candidate datasets.

4.2.4 Integrate

Selected datasets serve as input for the integration phase to onboard the external data into the company's systems and to document and prepare them for further use. This phase's activities begin by accessing the external data via a proposed interface. Along with an overview of the target integration system, the use case requirements help when choosing the appropriate access interface (e.g., download or API). Integration efforts are a challenging endeavor because of the external datasets' heterogeneity. The onboarding activities include thorough documentation of selected external datasets, which provides complete metadata information about the sources and their contents – specifically, the attributes. Our approach relies on the use of knowledge graphs to ensure the business concepts in external data correspond to the internal ones and can proceed by mapping both in a common data model. Knowledge graphs are considered an appropriate way of integrating heterogeneous datasets (Bizer et al., 2009; Paulheim, 2016) and have proved to be effective in the context of open data (Krasikov et al., 2021). In case there are multiple external datasets with semantically corresponding attributes, these sources should be combined by using a common data model. Newly onboarded external data requires ownership attributed to a dedicated data owner. The phase ends with M4 once external datasets have been integrated.

4.2.5 Manage and use

When the external datasets have been integrated, it is crucial to ensure their successful use. Our findings show that, following the integration of external data into the internal systems, its management should not diverge from the process established for internal data. This implies that external data is considered to be internal once it has been fully onboarded (see subsection 4.2.4). Nonetheless, provided the similarities between internal and external data, upon the integration of the latter, attention should be paid to the aspects that deviate. External data may need to receive updates from the original source and, thus, the update cycles should be managed by establishing datasets' versions and onboarding new data accordingly (see subsection 4.2.4). During its use, user feedback is collected to maintain the quality of the external data and identify improvement opportunities. Two scenarios were identified regarding the use of external data: Variant 5a – end date for the use of external data is anticipated; Variant 5b – continuous use of external data with no retirement planned. In case a decision to terminate the use of external data is taken and M5 is reached, the next phase will be embraced. The use of external data will continue until a retirement decision is taken.

4.2.6 Retire

The final phase aims to formalize the end-of-life of external data. The following variants summarize our findings of the possible scenarios in this phase: Variant 6a – the use case and the dataset are retired; Variant 6b – the use case remains, but new datasets are needed; Variant 6c – the dataset remains, but a new use case is needed.

Variant 6a assumes the complete termination of the process, where the use of external data is completely discontinued. Archiving the downloaded external data is a possibility when there are no contracts associated with its use. Conversely, it is important to monitor the termination terms of contracts/subscriptions in order not to bear the costs for the non-used data. Variant 6b implies that the process goes back to the screening phase (see subsection 4.2.2) and the search for new datasets begins. By contrast, if the use case is retired (6c) but the datasets remain in the internal systems, the process restarts with the definition of a new use case in the initial phase, respecting the possible Variant 1b.

4.3 Roles

Our findings, as well as the insights from the practitioners' reports, show that assigning the dedicated organizational roles along the process is essential for sourcing and managing external data (Aaser & McElhaney, 2021; Belissent, 2019; Explorium, 2021). Based on our learnings from the evaluation rounds and the insights from the company-specific instantiation of the process, we have identified that most of the tasks can be assumed by already existing roles within the company. We learned that this encompasses not only the data governance roles but also the associated functions such as procurement and compliance. For instance, the contracting process and negotiations with external data providers would rely on the procuring officer's efforts, while assessment would be the main task of a data steward, who ensures the data meets the desired standards of quality and is accessible for use within the company (van Donge et al., 2020). However, given the new activities related to the searching and screening of the relevant external data sources, we have identified an emerging role of a **data hunter**, who also acts as a domain expert. A data hunter's main responsibility is to find and review external data to ensure its fit within the defined use case (Aaser & McElhaney, 2021; Belissent, 2019). Therefore, they accumulate the expertise to manage external datasets upon its integration and can also serve as a company-internal single point of contact. Table 17 summarizes the roles for external data sourcing and management based on the (academic and practitioner) literature, expert input, and insights from the company-specific instantiation of the process.

Role	Responsibilities for core activities in the process
Requestor (data/business)	<ul style="list-style-type: none"> - Submits the request for external data - Uses the external data for selected use cases
Data analyst & business analyst	<ul style="list-style-type: none"> - Defines business and data requirements for future use cases - Identifies new use cases for external data, develops proofs-of-concept, and conducts analytics
Data architect/engineer	<ul style="list-style-type: none"> - Defines system requirements for external data integration
Data hunter/external data expert	<ul style="list-style-type: none"> - Partners with business to find, review, and manage external data (Aaser & McElhaney, 2021) - Monitors external data sources for relevant data to enable use cases - Acts as company-internal single point of contact for sourcing and managing external data - Provides first-level support internally, acts as an internal consultant for external data-related topics
Data steward	<ul style="list-style-type: none"> - Assesses the quality and fit of external datasets
Procurement	<ul style="list-style-type: none"> - Negotiates contracts and SLAs, analyzes the pricing conditions and contract termination
Compliance officer	<ul style="list-style-type: none"> - Ensures that external data is collected with appropriate permissions (contract and license agreements) and used in accordance with the applicable data laws and internal policies

Table 17. Roles involved in external data sourcing and managing

Our findings help to distinguish two configurations for the roles in the reference process. In the first and most common role model, existing roles undertake new activities proposed by the process. By contrast, the second model assumes that new tasks are taken over by the emerging role of data hunter or external data expert. Therefore, we conclude that new tasks relating to external data sourcing and managing can be either delegated to a new role or taken on by existing roles.

5 Discussion

While there are high expectations on external data's potential to fill gaps for reasonable decision-making and value creation, up until now there is no systematic approach to external data sourcing. Previous studies on conventional, crowd-based, and cloud-based data sourcing provide insights on generic steps of data sourcing, e.g., data demand, data acquisition, and data application (Wang et al., 2020). Studies dedicated to external data address relevant data sourcing activities, like identifying and evaluating potential data sources (Belissent, 2019; Strand & Carlsson, 2008). However, they do not provide reference processes but rather address external data activities from a data role perspective, e.g., the range of tasks for a data hunter (Belissent, 2019). The major advantages of the reference process that we have developed for sourcing and managing external data are (1) its ability to guide enterprises in their sourcing activities in a systematic way with clear milestones, and (2) its foundation on essential design and evaluations principles:

(1) *Guidance*: The reference model addresses the challenges associated with external data, such as the uncertainty of external datasets quality. Unexplored external datasets require a more thorough assessment than with traditional quality metrics (Zhang et al., 2019). Since external data creators and publishers are detached from their users (i.e., enterprises), the latter have limited knowledge about the data's characteristics and underlying quality. For the case of repurposed data, it is essential to adopt an approach that provides multiple perspectives on the sourced data (Krasikov et al., 2021).

(2) *Design principles*: An important design principle is to ensure that the obtained contents of the sourced external data are clear, and similar (or even the same) data has not already been integrated within the company. This principle builds on the transparency aspect of the external data (e.g., in terms of its provenance and trustworthiness) as one of the fuels for the digital economy. Having clear documentation of candidates and existing external data sources helps adhere to this principle. Since data cataloging activities are often seen as volunteer efforts (Jarvenpaa & Markus, 2020), this principle helps enforce their importance, particularly when it comes to the sourcing of new, unknown data within the internal systems. Another design principle is that at the handover, "reliable and relevant information is clearly communicated" (Maher et al., 2013). We addressed this design principle with the proposed milestones which present well-defined decision points that allow the involved parties to effectively communicate based on standardized performance criteria. Furthermore, to overcome the shortcomings of linear phase models, the reference process offers flexibility and process variability by formulating

11 possible variations of how the activities in the phases can be executed. They are also positioned as entry points, considering the current situation of the sourcing activities within the company. “Flexibility is an important characteristic of a reference process because it describes the ease with which a reference process accommodates and adapts to changes of the process requirements” (Wilmsen et al., 2020).

6 Conclusion

Although the use of external data is not new and has been mentioned in the enterprise context since the late 1990s (Čas & Meier, 1999), its sourcing is often associated with simply “getting the data” without any further specifications on how exactly this can be accomplished. In scientific literature, however, a systematic approach to sourcing and managing external data was lacking until now.

Accordingly, we propose a reference process for sourcing and managing external data that guides enterprises through unknown strides and that is methodologically well founded on design science research. In our industry–research collaboration, we leverage design science research to build a reference process that supports the external data sourcing and managing activities in the enterprise setting. Our reference process comprises six core phases which are described by the means inputs, activities and process variations, techniques, roles, and milestones.

To the best of our knowledge, this is the first study that addresses external data sourcing from a design science research perspective. This means that we do not only focus on the necessity of performing certain activities, but develop and refine a reference process in iterative design cycles. We managed to win four companies that are experienced and advanced in the field of data management to evaluate the reference process. Their feedback has been favorable with regard to the process’s relevance, understandability, and usefulness.

From an academic perspective, our findings synthesize and expand the scarce body of knowledge on data sourcing by proposing a reference process as a generic procedure (Goeken, 2011). We notice that the proposed reference process shows commonalities with strategic sourcing processes, but also uncovers data sourcing specificities. The latter include, for instance, the importance of semantics and concept mapping to integrate external data.

From the practitioners’ perspective, the designed reference process aims to solve the increasingly relevant organizational challenges and contributes to the professionalization of external data sourcing. This enables a shift from ad-hoc sourcing practices to a well-defined approach for sourcing and managing external data.

Since we are among the first to explore this area, certainly there are limitations, and further research is needed to gain a deeper understanding of the domain. Although our findings were well-perceived throughout the instantiations and the focus group discussions, no large-scale demonstrations or evaluations have been conducted yet. Thus, we foresee future research activities to apply the reference process in diverse use cases and enterprise contexts, which would help generalize our findings and identify situational configurations. While our study focused on a reference process, it provides some first insights on the emerging roles in the context of external data sourcing and management. Studying both aspects would allow to gain a broader perspective on external data governance mechanisms. Another promising avenue for future research is an opportunity to explore and discuss the peculiar nature of digital data as semantic resources, drawing upon emerging literature on the topic (Aaltonen et al., 2021; Aaltonen & Penttinen, 2021).

The reference process for sourcing and managing external data can support data hunters and decision makers in organizing their activities, but it is not a foregone conclusion, thus specific characteristics of enterprises' data foundation must always be included.

7 References

- Aaltonen, A., Alaimo, C., & Kallinikos, J. (2021). The Making of Data Commodities: Data Analytics as an Embedded Process. *Journal of Management Information Systems*, 38(2), 401–429.
- Aaltonen, A., & Penttinen, E. (2021). *What Makes Data Possible? A Sociotechnical View on Structured Data Innovations*. Proceedings of the 54th Hawaii International Conference on System Sciences.
- Aaser, M., & McElhaney, D. (2021). *Harnessing the Power of External Data*. McKinsey. <https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/harnessing-the-power-of-external-data>
- Amsterdamer, Y., & Milo, T. (2015). Foundations of Crowd Data Sourcing. *ACM Special Interest Group on Management of Data (SIGMOD) Record*, 43(4), 5–14.
- Arndt, D., & Gersten, W. (2001). External Data Selection for Data Mining in Direct Marketing. *Proceedings of the Sixth International Conference on Information Quality*, 44–61.
- Baecke, P., & Van den Poel, D. (2011). Data Augmentation by Predicting Spending Pleasure Using Commercially Available External Data. *Journal of Intelligent Information Systems*, 36(3), 367–383.
- Baud, N., Frachot, A., & Roncalli, T. (2002). Internal Data, External Data and Consortium Data—How to Mix Them for Measuring Operational Risk. *SSRN Electronic Journal*, 1–18.
- Becker, J., Delfmann, P., & Knackstedt, R. (2007). Adaptive Reference Modeling: Integrating Configurative and Generic Adaptation Techniques for Information Models. In J. Becker & P. Delfmann (Eds.), *Reference Modeling* (pp. 27–58). Physica-Verlag HD.
- Belissent, J. (2019). *The Insights Professional's Guide to External Data Sourcing*. Forrester Research. <https://www.forrester.com/report/The-Insights-Professionals-Guide-To-External-Data-Sourcing/RES139331>
- Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data—The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3), 1–22.
- Čas, K., & Meier, M. (1999). Integration of Internal and External Data for Marketing Management. In *Evolution and Challenges in System Development* (pp. 489–503). Springer.
- Clark, T. D., Zmud, R. W., & Mccray, G. E. (1995). The Outsourcing of Information Services: Transforming the Nature of Business in the Information Industry. *Journal of Information Technology*, 10(4), 221–237.
- Davenport, T. H., Evgeniou, T., & Redman, T. C. (2021). Your Data Supply Chains Are Probably a Mess. Here's How to Fix Them. *Harvard Business Review*. <https://hbr.org/2021/06/data-management-is-a-supply-chain-problem>
- Deutch, D., & Milo, T. (2012). Mob Data Sourcing. *Proceedings of the 2012 International Conference on Management of Data*, 581–583.
- Dun & Bradstreet. (2018). *Global Family Tree Report*. <https://docs.dnb.com/onboard/en-GB/Business/Reports/global-family-tree-report>
- Explorium. (2021). *State of External Data Acquisition*. <https://www.explorium.ai/resource/explorium-2021state-of-external-data-acquisition/>
- Goeken, M. (2011). Towards an Evidence-based Research Approach in Information Systems. *Proceedings of the 32nd International Conference on Information Systems*.
- Gregor, S. (2006). The Nature of Theory in Information Systems. *MIS Quarterly*, 30(3), 611–642.
- Hopf, K. (2019). *Predictive Analytics for Energy Efficiency and Energy Retailing*. PhD thesis, University of Bamberg.
- Jarvenpaa, S. L., & Markus, M. L. (2020). Data Sourcing and Data Partnerships: Opportunities for IS Sourcing Research. In *Information Systems Outsourcing* (5th ed., pp. 61–79). Springer.
- Könning, M., Westner, M., & Strahringer, S. (2019). A Systematic Review of Recent Developments in IT Outsourcing Research. *Information Systems Management*, 36(1), 78–96.
- Kotlarsky, J., Oshri, I., Dibbern, J., & Mani, D. (2018). *MIS Quarterly Research Curation on IS Sourcing*. <https://www.misqresearchcurations.org/blog/2018/6/26/is-sourcing>
- Krasikov, P., Legner, C., & Eurich, M. (2021). Sourcing the Right Open Data: A Design Science Research Approach for the Enterprise Context. In *The Next Wave of Sociotechnical Design* (Vol. 12807, pp. 313–327). Springer.
- Lacity, M. C., Khan, S. A., & Yan, A. (2016). Review of the Empirical Business Services Sourcing Literature: An Update and Future Directions. *Journal of Information Technology*, 31(3), 269–328.
- Lacity, M. C., Khan, S., Yan, A., & Willcocks, L. P. (2010). A Review of the IT Outsourcing Empirical Literature and Future Research Directions. *Journal of Information Technology*, 25(4), 395–433.
- Legner, C., Pentek, T., & Otto, B. (2020). Accumulating Design Knowledge with Reference Models: Insights from 12 Years' Research into Data Management. *Journal of the Association for Information Systems*, 21(3), 735–770.
- Lukyanenko, R., Parsons, J., Wiersma, Y., & Maddah, M. (2019). Expecting the Unexpected: Effects of Data Collection Design Choices on the Quality of Crowdsourced User-Generated Content. *MIS Quarterly*, 43(2), 634–647.
- Lyytinen, K., Sørensen, C., & Tilson, D. (2017). Generativity in Digital Infrastructures: A Research Note. In *The Routledge Companion to Management Information Systems* (1st ed., pp. 253–275). Routledge.
- Maher, B., Drachslar, H., Kalz, M., Hoare, C., Sorensen, H., Lezcano, L., Henn, P., & Specht, M. (2013). Use of Mobile Applications for Hospital Discharge Letters – Improving Handover at Point of Practice. *International Journal of Mobile and Blended Learning*, 5(4), 19–42.

- March, S. T., & Smith, G. F. (1995). Design and Natural Science Research on Information Technology. *Decision Support Systems*, 15(4), 251–266.
- McCarthy, S., Rowan, W., Lynch, L., & Fitzgerald, C. (2020). Blended Stakeholder Participation for Responsible Information Systems Research. *Communications of the Association for Information Systems*, 47, 716–742.
- Muhic, M., & Johansson, B. (2014). Cloud Sourcing – Next Generation Outsourcing? *Procedia Technology*, 16, 553–561.
- Nevo, D., & Kotlarsky, J. (2020). Crowdsourcing as a Strategic IS Sourcing Phenomenon: Critical Review and Insights for Future Research. *The Journal of Strategic Information Systems*, 29(4), 101593.
- Nordin, F., & Henrik, A. (2008). Business Service Sourcing: A Literature Review and Agenda for Future Research. *International Journal of Integrated Supply Management*, 4(3/4), 378–405.
- Oshri, I., Kotlarsky, J., & Willcocks, L. P. (2015). *The Handbook of Global Outsourcing and Offshoring* (3rd ed.). Palgrave Macmillan.
- Paulheim, H. (2016). Knowledge Graph Refinement. *Semantic Web*, 8(3), 489–508.
- Peffer, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, 24(3), 45–77.
- Prat, N., Comyn-Wattiau, I., & Akoka, J. (2015). A Taxonomy of Evaluation Methods for Information Systems Artifacts. *Journal of Management Information Systems*, 32(3), 229–267.
- Rafati, L., & Poels, G. (2015). Towards Model-Based Strategic Sourcing. In *Achieving Success and Innovation in Global Sourcing: Perspectives and Practices* (Vol. 236, pp. 29–51). Springer.
- Ribas, I., Lusa, A., & Corominas, A. (2021). Multi-Step Process for Selecting Strategic Sourcing Options When Designing Supply Chains. *Journal of Industrial Engineering and Management*, 14(3), 477–495.
- Roeder, J., Muntermann, J., & Kneib, T. (2020). Towards a Taxonomy of Data Heterogeneity. *Proceedings of the 15th International Conference on Wirtschaftsinformatik*, 293–308.
- Satish, L., & Yusof, N. (2017). A Review: Big Data Analytics for Enhanced Customer Experiences with Crowd Sourcing. In *Procedia Computer Science* (Vol. 116, pp. 274–283).
- Schatsky, D., Camhi, J., & Muraskin, C. (2019). *Data Ecosystems: How Third-Party Information Can Enhance Data Analytics*. Deloitte. https://www2.deloitte.com/content/dam/insights/us/articles/4603_Data-ecosystems/DI_Data-ecosystems.pdf
- Schneider, C. O., Bremen, P., Schönsleben, P., & Alard, R. (2013). Transaction Cost Economics in Global Sourcing: Assessing Regional Differences and Implications for Performance. *International Journal of Production Economics*, 141(1), 243–254.
- Schneider, S., & Sunyaev, A. (2016). Determinant Factors of Cloud-Sourcing Decisions: Reflecting on the IT Outsourcing Literature in the Era of Cloud Computing. *Journal of Information Technology*, 31(1), 1–31.
- Sesay, A., & Ramirez, R. (2016). Theorizing the IT Governance Role in IT Sourcing Research. *Proceedings of the 22nd Americas Conference on Information Systems*.
- Strand, M., & Carlsson, S. A. (2008). Provision of External Data for DSS, BI, and DW by Syndicate Data Suppliers. *Proceedings of the 2008 Conference on Collaborative Decision Making: Perspectives and Challenges*, 245–256.
- Strand, M., & Syberfeldt, A. (2020). Using External Data in a BI Solution to Optimise Waste Management. *Journal of Decision Systems*, 29(1), 53–68.
- Sun, Z., Di, L., Fang, H., Guo, L., Tan, X., Jiang, L., & Chen, Z. (2021). Agro-Geoinformatics Data Sources and Sourcing. In *Agro-Geoinformatics: Theory and Practice* (pp. 41–66). Springer.
- van Donge, W., Bharosa, N., & Janssen, M. F. W. H. A. (2020). Future Government Data Strategies: Data-Driven Enterprise or Data Steward?: Exploring Definitions and Challenges for the Government as Data Enterprise. *Proceedings of the 21st Annual International Conference on Digital Government Research*, 196–204.
- Venable, J., Pries-Heje, J., & Baskerville, R. (2012). A Comprehensive Framework for Evaluation in Design Science Research. In *Design Science Research in Information Systems. Advances in Theory and Practice* (Vol. 7286, pp. 423–438). Springer.
- Wang, W. M., Preidel, M., Fachbach, B., & Stark, R. (2020). Towards a Reference Model for Knowledge Driven Data Provision Processes. In *Boosting Collaborative Networks 4.0* (Vol. 598, pp. 123–132). Springer.
- Watjatrakul, B. (2005). Determinants of IS Sourcing Decisions: A Comparative Study of Transaction Cost Theory Versus the Resource-Based View. *The Journal of Strategic Information Systems*, 14(4), 389–415.
- Wilmsen, M., Gericke, K., Jäckle, M., & Albers, A. (2020). Method for the Identification of Requirements for Designing Reference Processes. *Proceedings of the Design Society: DESIGN Conference*, 1, 1175–1184.
- Winter, R., & Schelp, J. (2006). Reference Modeling and Method Construction: A Design Science Perspective. *Proceedings of the 2006 ACM Symposium on Applied Computing*, 1561–1562.
- Zhang, R., Indulska, M., & Sadiq, S. (2019). Discovering Data Quality Problems: The Case of Repurposed Data. *Business & Information Systems Engineering*, 61(5), 575–593.
- Zuiderwijk, A., Janssen, M., Poulis, K., & van de Kaa, G. (2015). Open Data for Competitive Advantage. *Proceedings of the 16th Annual International Conference on Digital Government Research*, 79–88.

Essay 3

Open Data in the Enterprise Context: Assessing Open Corporate Data's Readiness for Use

Pavel Krasikov, Timo Obrecht, Markus Eurich, and Christine Legner

Faculty of Business and Economics (HEC), University of Lausanne, Switzerland

*First version published in the Proceedings of the 9th International Conference on Data Science,
Technology and Applications (DATA), 2020*

*Extended version published in the Communications in Computer and Information Science (CCIS)
book series, 2021*

Abstract: *Open data initiatives have long focused on motivating governmental bodies to open up their data. Although governments and other organizations make their data increasingly available, open data consumers are reluctant or experience difficulties with using open data. Prior studies have therefore set the focus on open data portals and open data quality, but only few have examined enterprises as consumers of open data. To close this gap, we aim at assessing whether open data is ready for use by enterprises. We focus our efforts on open corporate data, i.e., data on companies provided by business registers, which has confirmed reuse potential. Our assessment of 30 business registers confirms that the heterogeneity of access, licensing, publishing conditions, and content in open corporate datasets hinder their reuse in a business context. Only half of analyzed registers provide companies' full legal addresses, only 20% mention their complete organizational information, while contact details are fully available in 13% of all the cases. We find that open data's readiness for use from an enterprise perspective is highly dependent on the concrete use case.*

Keywords: Open data, Corporate registers, Open corporate data, Usability, Data quality, Open data assessment

Table of contents

- 1 Introduction..... 106
- 2 Related work107
 - 2.1 Barriers to open data adoption.....107
 - 2.2 Open data assessment 109
 - 2.3 Research gap 110
- 3 Methodology..... 111
- 4 Findings..... 114
 - 4.1 Data sources identification 114
 - 4.2 Metadata analysis.....115
 - 4.3 Schema analysis..... 116
 - 4.4 Ready for use assessment117
- 5 Conclusion121
- 6 References 123
- Appendix 1..... 125
- Appendix 2126

List of figures

Figure 6. Research process	112
Figure 7. Schema analysis: presence of attributes across the analyzed registers.....	116
Figure 8. Schema analysis: presence of attributes per corporate register	117
Figure 9. Metadata analysis of corporate registers 1 to 10.....	126
Figure 10. Metadata analysis of corporate registers 11 to 20	126
Figure 11. Metadata analysis of corporate registers 21 to 30.....	127

List of tables

Table 18. Barriers to open data adoption.....	109
Table 19. Open data assessments	110
Table 20. Focus group composition	112
Table 21. Analyzed corporate registers with GLEIF identifier.....	115
Table 22. Analysis of master data management use cases.....	118
Table 23. Analysis of fraud prevention use case	119
Table 24. Analysis of intelligence and analytics use case	119
Table 25. Analysis of marketing use case.....	120
Table 26. Definition of attributes	125

1 Introduction

Open data can be defined as “data that is freely available, and can be used as well as republished by everyone without restrictions from copyright or patents” (Braunschweig et al., 2012). It holds great business potential, with global economic value estimated as from \$3.2 to \$5.4 trillion annually (Manyika et al., 2013), as well as forecasted cost savings of 1.7 billion EUR for the EU28+ countries (European Commission et al., 2015). Open data initiatives have long focused on motivating governments to open their data (Zuiderwijk et al., 2012). However, although the number of open datasets is growing steadily, their adoption is lagging behind (Publications Office of the EU, 2020). In the first wave, application developers were the main users of open data, achieving only modest success (Bizer et al., 2009). In the current second wave, authorities as well as national and European initiatives are pushing open data’s wider adoption and using it to create added value (Puha et al., 2018). It is widely believed that multiple industry sectors could significantly benefit from open data, among them transportation, consumer products, electricity, oil and gas, healthcare, consumer finance, agriculture, urban development, and the social sector (Davies et al., 2019; Deloitte Analytics, 2012; Dinter & Kollwitz, 2016; Manyika et al., 2013; Publications Office of the EU, 2020). Despite the significant business potential, enterprises are far from leveraging the available open data resources and most of them are reluctant to even try (Davies et al., 2019; Oliveira et al., 2016). This is due to a lack of transparency, unknown quality, and unclear licensing unsettling challenges (Janssen et al., 2012; Martin et al., 2013).

In this study, we assess whether open data is ready for use in the enterprise context. We focus on one of the most important segments of open government data: open corporate data (OCD). OCD can be defined as data on companies that corporate registers (also known as business registers), in keeping with local laws, usually collect. This data is transparent and interoperable, and has a confirmed reuse potential (Varytimou et al., 2015). Our study extends our earlier conference paper (Krasikov et al., 2020) which provided an initial analysis of OCD provided by business registers from different countries. It addresses the following research questions:

To which extent is open corporate data ready for use by enterprises?

Does open corporate data satisfy the requirements of typical enterprise use cases?

Compared to our conference publication (Krasikov et al., 2020), we improve and refine the use case-driven analysis of OCD. The suggested approach (see Figure 6) extends beyond metadata and schema analysis and incorporates the “ready for use” assessment. This additional step (see subsection 4.4) comprises the mapping of required business concepts with OCD attributes and

is demonstrated in four typical use cases. In addition, we revisited the originally considered business registers, updated the metadata and content analysis as of October 2020, and added 10 new corporate registers.

In total, we analyze data from 30 open corporate registers: first, by assessing the provided metadata and, second, by examining the content of these corporate registers. To assess whether open corporate data is ready for use, we compare the datasets' content with the common data objects that typical use cases require. Our findings confirm that the heterogeneity of access, licensing, publishing conditions, and content in open corporate datasets hinder their reuse in a business context. In addition, our study shows that only half of analyzed registers provide companies' full legal addresses, only 20% mention their complete organizational information, while contact details are fully available in 13% of all the cases. For four typical use cases (master data management, fraud prevention, intelligence and analytics, and marketing), we conclude that open corporate data has only limited use due to its lacking coverage of relevant business concepts. Our study thereby underlines shortcomings in business registers, but also draws attention to the need for domain-specific semantic models that make open data more usable for enterprises.

The remainder of the paper is organized as follows: In section 2, we examine relevant literature on open data adoption barriers and assessment techniques, which clarifies the research gap. In the section that follows, we explain our research methodology. In section 4 we thoroughly describe this study's results. Finally, we present our concluding remarks, discuss the study's limitations, and provide suggestions for future research.

2 Related work

While governments and other organizations make their data increasingly available, open data consumers are reluctant or experience difficulties with using open data. In this section, we will review research on the barriers to open data adoption from both providers' and consumers' perspective and analyze various assessment methods, which have been proposed in prior literature.

2.1 Barriers to open data adoption

Prior studies on the barriers to open data adoption differentiate between open data consumption and supply (see Table 1). Although the barriers are associated with either consumption or supply, there is a strong interdependency between the two: the way the data is published impacts how

it is used (Zuiderwijk et al., 2012, fig. 1). Consequently, most studies investigate both consumption and supply.

When it comes to data provisioning, these studies identify several common issues: the risk of excessive costs, an unclear purpose, as well as litigation and differing licensing standards and documentation complicating open data suppliers' release process (Martin et al., 2013; Barry & Bannister, 2014; Conradie & Choenni, 2014; Beno et al., 2017). Studies addressing consumption barriers emphasize that the setbacks are not strictly technical (Beno et al., 2017; Martin et al., 2013; Zuiderwijk et al., 2012), but also concern the broader context of data use. The absence of information describing an open dataset is often associated with poor metadata documentation (Zuiderwijk et al., 2012). The latter generally refers to technical barriers, demonstrating the interdependence of the impediments' consumption and supply sides. In addition, a lack of understanding of the contents and insufficient domain knowledge commonly hinder open data use (Beno et al., 2017; Janssen et al., 2012; Zuiderwijk et al., 2012). As underlined by (Krasikov et al., 2020), three challenges prevail in using open data: first, there is a lack of transparency about datasets' availability and their usefulness for the end user (Janssen et al., 2012). Second, open datasets' heterogeneity in terms of licensing conditions, available formats, and access to information complicates the integration efforts (Martin et al., 2013). Third, the quality of open data remains unknown and uncertain in terms of typical assessment criteria (Zuiderwijk et al., 2012).

Finally, it is worthwhile noting that many of the existing studies do not consider any specific use context, and only two studies examined enterprises as consumers of open data. This underpins the lack of research on open data use in the enterprise context.

Source and topic	Method	Adoption barriers	Open data
(Janssen et al., 2012) Gap between the benefits of and barriers to open data adoption	Group session (n=9), findings were discussed during interviews (n= 14)	6 categories: institutional, task complexity, use and participation, legislation, information quality, technical. Categories are exemplified by a total of 57 examples of barriers	Generic consumption Supply
(Zuiderwijk et al., 2012) Open data users' perspective on identified impediments	Literature review (n=37) Interviews (n=6) Workshops (n=4)	A total of 118 socio-technical impediments in 3 categories: data access, data use, and data deposition. 10 sub-categories: availability and access, findability, usability, understandability, quality, linking and combining data, comparability and compatibility, metadata, interaction with data provider, and opening and uploading	Generic consumption Supply

Source and topic	Method	Adoption barriers	Open data
(Martin et al., 2013) Risks for re-users of public data differ from those for open data providers	Analysis of open data platforms (n=3)	Typology of barriers comprising 7 categories: governance, economic issues, licenses and legal frameworks, data characteristics, metadata, access, and skills	Business consumption Supply
(Conradie & Choenni, 2014) Release processes of government open data	Participatory action research: Exploratory workshop (n=5). Questionnaire answered by a consortium (n=14). Questionnaire answered by other civil servants (n=50). In-depth interviews (n=18). Workshop with data users (n=8). Plenary session discussion (n=21). Follow-up meeting with decision makers (n=2). Experiences with data release (n=4)	4 categories of barriers: fear of false conclusions, financial effects, opaque ownership and unknown data locations, and priority	Supply
(Barry & Bannister, 2014) Implications of opening up the data	Case studies (n=2), inductive approach to the analysis of collected data	6 types of barriers: economic, technical, cultural, legal, administrative, and task related. A total of 20 barriers to open data's release	Supply
(Beno et al., 2017) Practitioners using and providing open data in Austria	Literature review (n=17) Survey (n=110)	3 major groups: user specific, provider specific, and both users and providers with a total of 54 barriers	Consumption by enterprises, academia, and public sector Supply

Table 18. Barriers to open data adoption

2.2 Open data assessment

Since open data portals play an important role in publishing open data, researchers have set their focus on their assessment. Table 2 summarizes the scope of prior studies and the ways they assessed open data. It sheds light on two crucial aspects in open data assessment studies: (1) whether the unit of analysis was the metadata or the dataset content as well, and (2) the methods used. We find that prior research almost exclusively focuses on the metadata quality. Although “poor data quality can be widespread, and potentially hamper an efficient reuse of open data” (2016), only three studies analyze the contents of the underlying datasets. Interestingly, these studies build on generic quality assessment methods according to typical data quality dimensions, such as completeness, accuracy, or timeliness. They neither consider specific data requirements nor the use contexts, although data quality is commonly defined as “fitness for use” from the data consumers' perspective (Wang & Strong, 1996). This means that the reviewed literature largely ignores the actual user's perspective and the data domain knowledge, which has found to be crucial for overcoming the barriers (section 2.1). As a final point, open data's

usefulness is only addressed by Osagie et al. (2017), who take a very specific focus on the usability of open data platforms' features for specific use cases.

Source	Scope	Unit of analysis	Assessment approach
(Bogdanović-Dinić et al., 2014)	7 open data portals	Metadata	"Data openness score" based on eight open data principles (Open Government Working Group, 2007)
(Reiche et al., 2014)	10 open government data portals	Metadata	Ranking of open data repositories with the average score computed by means of quality metrics
(Umbrich et al., 2015)	82 CKAN portals	Metadata	Open data quality and monitoring assessment framework with 6 quality dimensions
(Neumaier et al., 2016)	260 open data portals	Metadata	Metadata quality assessment framework
(Vetrò et al., 2016)	11 datasets	Metadata and dataset	Quality framework supported by data quality models from the literature, 6 dimensions and 14 metrics
(Máchová & Lněnička, 2017)	67 open data portals	Metadata	Benchmarking framework for evaluating open data portals' quality
(Welle Donker & Van Loenen, 2017)	20 "most wanted" datasets in Netherlands	Metadata	Holistic open data assessment framework with 3 main levels: open data supply, open data governance, and open data user characteristics.
(Osagie et al., 2017)	5 datasets	Platform features	Usability evaluation with ROUTE-TO-PA and QUIN criteria. (12 usability criteria)
(Bicevskis et al., 2018)	4 company registers for 11 attributes	Dataset	Three-part data quality model (syntax analysis): definition of a data object, data object quality specifications, and implementation
(Kubler et al., 2018)	More than 250 open data portals	Metadata	Open data portal quality (ODPQ) framework with 17 quality dimensions
(Stróżyna et al., 2018)	59 data sources	Metadata	Quality-based selection, assessment, and retrieval method. Attribution of quality scores based on "ranking type Delphi" and 6 quality dimensions
(Zhang et al., 2019)	20 datasets	Metadata and dataset	Design science research and a systematic approach to repurposed datasets' quality using the LANG approach and according to 10 dimensions

Table 19. Open data assessments

2.3 Research gap

One of the least addressed barriers in the open data landscape is the "lack of insight into the user's perspective" (Janssen et al., 2012). This implies understanding the particularities of open data access, publishing, licensing, and content, as well as the extent to which they meet the requirements in a specific use context and business scenario. Existing efforts study barriers mostly on the platform level, rather than on the dataset level (Osagie et al., 2017) or refer to open

data supply and its underlying technical impediments evoking users' behavioral intentions (Weerakkody et al., 2017).

Furthermore, the literature does not specifically cover open data's use in the business context. For instance, governmental directives to open basic data about companies (European Parliament, 2012) motivate the competent authorities to make this data available. However, this does not necessarily mean that the data is also usable (Varytimou et al., 2015). Existing literature often restricts the user's perspective to data availability (the way data is proposed and can be consumed) by considering usability purely in terms of technical specifications (Osagie et al., 2017; Weerakkody et al., 2017), such as data format and open data portal's underlying software. We argue for taking a user-centric perspective based on Welle Donker et al's (2017) definition of open data's usability as "usable for the intended purpose of the user." In fact, being manageable is one of the indicators that the authors introduce in the same work, which implies that a "user should be able to use it (open data) with available resources and for the goal the user had in mind" (Welle Donker & Van Loenen, 2017).

The abovementioned gaps motivate our research aimed at answering the question whether open data is ready for use by enterprises.

3 Methodology

In this study, we address the identified research gap by assessing whether open data is ready for use in a specific domain and for the enterprise context. We selected open corporate data (OCD), which is an important segment of open government data. OCD can be defined as data on companies that business registers, in keeping with local laws, usually collect. The resulting data is not only valuable for public authorities, but its high potential for reuse in a business setting has been emphasized by practitioners and researchers (Koznov et al., 2016; Varytimou et al., 2015).

Our research process comprised different research activity to assess open corporate data: a literature analysis to understand open data's current state and its adoption barriers; focus groups with practitioners to specify use cases in the enterprise context; the in-depth assessment of open corporate sources and datasets in the form of metadata and content analysis. Figure 6 summarizes the key phases of the research process, where the numbers refer to the corresponding sections in the results of this paper.

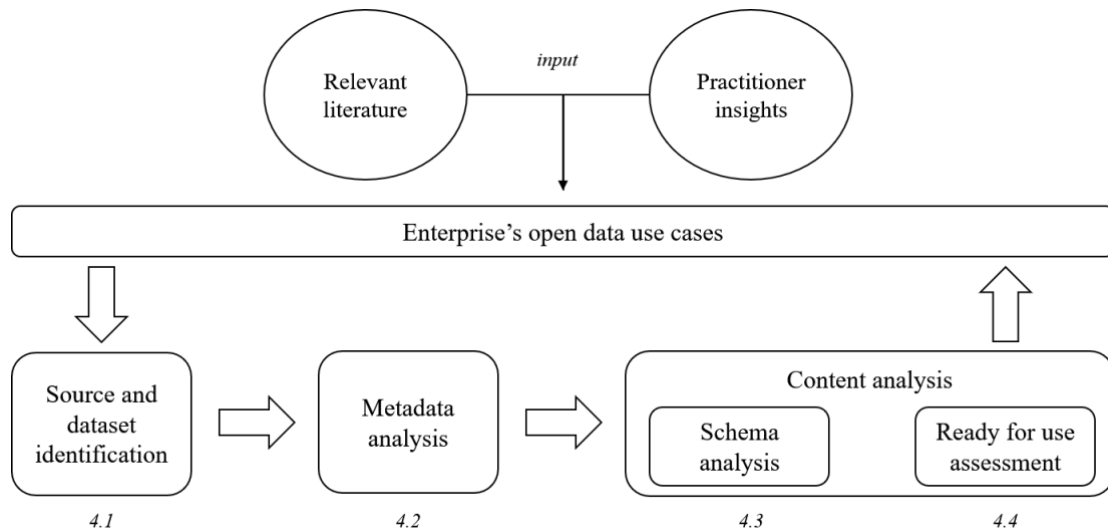


Figure 6. Research process

Enterprise’s Open Data Use Cases. To discuss and analyze the use context of OCD in enterprises, we formed a focus group (Bryman & Bell, 2007, p. 511; Creswell, 2009, p. 181) with seven Swiss-based data management experts representing transportation, consumption goods, and telecommunication industries (see Table 3). This activity was a part of a broader analysis of open data use cases. All the participants were knowledgeable about open data use cases and had been involved in the generation and documentation processes of OCD scenarios. The focus group first met during a Web conference during which it defined three high level use cases based on a structured use case generation framework (Krasikov et al., 2019). Afterwards, the focus group met physically for a workshop that validated open corporate data use cases. Additional individual sessions were conducted with the same companies to refine the relevant use cases and obtain further insights. This activity resulted in four concrete use cases and the corresponding business concepts, which could potentially be sourced from OCD datasets (see section 4.4).

Company	Industry	Size	Key informants	OCD context
A	Public transportation and mobility infrastructure	Revenue: \$1B to \$50B Employees: ~35 000	Data architect, open data responsible	Data management and business processes improvement
B	Consumer goods	Revenue: \$50B to \$100B Employees: ~300 000	Leader: data management and business analytics	Enhancement of business processes, business intelligence and analytics
C	Telecommunication	Revenue: \$1B to \$50B Employees: ~20 000	Head of data quality management	Validation, enrichment, and deduplication of internal data

Table 20. Focus group composition

Data Sources and Datasets Identification. Since academic literature on OCD is scarce, we mainly used online sources as providers of insights (GLEIF, 2019; Global Open Data Index, 2015; OpenCorporates, 2020; OpenDataBarometer & World Wide Web Foundation, 2020; Wikipedia, 2019). To date, 737 official corporate registers from 224 countries have been recorded following the Global Legal Entity Identifier Foundation's (GLEIF) accreditation process (GLEIF, 2017). However, only a small number of these officially confirmed registers are available as open sources. In our analysis we considered 30 corporate registers provided by official government agencies that have their data available in full open access (Stróżyńska et al., 2018). We have extended and updated the previous list (Krasikov et al., 2020) by adding the leading countries in the open data initiatives in EU (Publications Office of the EU, 2020) and world leading economies with recognized open data initiatives (Global Open Data Index, 2015; OpenCorporates, 2020; OpenDataBarometer & World Wide Web Foundation, 2020). Our selection covers corporate registers of United States, Europe, and other countries, and considers different geographical granularity. Many registers claim their data to be provided with an open license, whereas the real access is restrained by registration, forms submissions, or even blocked by fees for downloads or API calls. Even though we initially wanted to consider registers that provide strictly full access to the data (such as bulk download or API), we realized, during the course of the analysis, that some claiming to have an open license only allow partial access to the datasets, for example, the Austrian, Belgian, Danish, Indian, Swiss etc. (see Table 4) business registers.

Metadata Analysis. As seen in section 2.2, most of the open data assessment methods focus on metadata. In fact, the primary insights into whether the desired data is usable or not are obtained through the metadata published at the source. We relied on previous literature (see Table 1) when dealing with corporate registers and collected five categories of open data information: its identification, access, licensing, publisher, and basic information about the underlying datasets' content (see Appendix 1, Source Information). Two researchers collected and reconciled the metadata of the selected 30 corporate registers (see Appendix 2).

Schema Analysis. Following the research process, a comprehensive content analysis of the corporate registers was undertaken to assess its readiness for use (see section 4.3). Two researchers conducted a bottom-up analysis to understand the similarities between the attributes that the registers provide. Based on the focus group participants' input, we examined the corporate registers to ascertain the presence of attributes related to the use cases' relevant business concepts. Moreover, we took existing efforts regarding the OCD semantics' standardization into consideration for this analysis. We derived 21 typically used attributes (see

Appendix 1, Content Information), based on the analysis of the selected business registers. Furthermore, we have assessed the presence of these attributes within the 30 selected registers.

Ready for Use Assessment. Finally, we investigate OCD's readiness for use in a business setting, by determining the presence of required business concepts within the analyzed corporate registers. We specifically analyzed four use cases analyzed in the first step, i.e., master data management, fraud prevention, intelligence and analytics, and marketing. For these scenarios, the participants of the focus group helped to identify relevant data objects, which we then compared to the business concepts typically found in corporate registers. Section 4.4 elaborates on our findings regarding the usability of OCD for the given usage scenarios.

4 Findings

This section summarizes our findings along the different steps in the research process, i.e., identification of relevant datasets provided by the corporate registers (4.1), their assessment in terms of metadata documentation (4.2), schema analysis (4.3), and ready for use assessment with presence of business concepts required for the identified use cases (4.4).

4.1 Data sources identification

Corporate registers are usually assigned to a country or an administrative area and cover local business entities that need to undergo a local registration procedure. Aggregated unofficial lists of existing company registers are available online per country (Wikipedia, 2019), although there is no assessment process that confirms this sources' accuracy. The abovementioned GLEIF has an attribution procedure by means of a legal entity identifier (LEI), and maintains a catalogue with accredited official business registers, which provides initial insights into the available OCD (GLEIF, 2019). The register's presence on this list does not guarantee that the provided data is open. For instance, the Austrian corporate register maintained by the Federal Ministry of Justice (Krasikov et al., 2020), is currently available only at a fee. For this reason, we have selected an open Austrian business register provided by a different publisher.

For our analysis, we have selected 30 sources, i.e., corporate registers covering United States, Europe, and other countries worldwide with advanced open data initiatives, as listed in Table 4.

Alaska Business Entity Register (RA000594)	Norway Register of Business Enterprises (RA000472)
Argentinian National Registry of Companies (RA000010)	Oregon Business Entity Register (RA000631)
Australian Business Register (RA000013)	Russian Register of Legal Entities (RA000499)
Bulgarian Commercial Register (RA000065)	Singapore ACRA Register (RA000523)
Canada Corporate Register (RA000072)	UK Companies House (RA000585)
Colorado Business Entity Register (RA000599)	Ukrainian State Register Service (RA000567)
Finnish Business Information System (RA000188)	Washington Business Entity Register (RA000641)
Florida Business Entity Register (RA000603)	Wyoming Business Entity Register (RA000644)
France Register of Companies (RA000189)	Austrian Corporate Register (RA000687)*
Iowa Business Entity Register (RA000606)	Central Belgium Company Database (RA000025)*
Ireland Companies Register (RA000402)	Cyprus Companies Section (RA000161)*
Japanese National Tax Agency (RA000413)	Danish Company Register CVR (RA000170)*
Latvian Register of Enterprises (RA000423)	Indian Business Register (RA000394)*
Moldova State Register of Legal Entities (RA000451)	New Zealand Company Register (RA000466)*
New York Business Entity Register (RA000628)	Swiss UID-Register (RA000548)*
<i>Corporate registers provide full access to the data, except for the ones marked with * (restricted access)</i>	

Table 21. Analyzed corporate registers with GLEIF identifier

4.2 Metadata analysis

The analysis of the collected metadata provides first insights into the sources. Appendix 2 (see Figures 4-6) summarize the metadata documentation for the business registers and present identification information regarding the relevant countries and GLEIF registry codes, which allows to identify the webpage for each dataset.

Metadata regarding access revealed some interesting insights. Registers, which provided bulk download option, most frequently relied on such machine readable and suitable for processing file formats as CSV, JSON, and XML. A growing number of registers (11) started to offer APIs as an access point to the data. Five registers required a login procedure in order to obtain a full access to the data, but still offering the possibility of a lookup service with limited access. Moreover, with the exception of one, all of the registers provided a free lookup service to query the register. In terms of licensing, vast majority of registers operated under an open license, a Creative Commons one or a national equivalent, whereas seven registers provided access to the data without any license specification. More than a half (16) of the business registers noted a publishing date, which was after 2013. The data's update frequency varied from daily or weekly to monthly or even yearly. Finally, these attributes' importance should not be underestimated

(Kampars et al., 2020) as they are an integral part of the enterprises' specific needs (see Section 4.4).

Metadata regarding the content reveals an important difference between the registers' sizes, which ranged from 82,902 to 21,059,740 entry points. Their geographical coverage explains this, as larger registers cover the national level of granularity (France, Australia, and UK), while smaller ones are restricted to states (US) or administrative areas. We also notice that the revisited registers are not static and continue to grow compared to our previous analysis (Krasikov et al., 2020). Almost all the registers are available in English, even though the country of origin has a different national language or more than one, which demonstrates the efforts taken to make data available for an international audience. While this information allows first insights into the data, we provide a thorough analysis of the contents and mapping to the identified use cases in the following sections.

4.3 Schema analysis

Upon the identification and documentation of the corporate registers, we analyzed the presence of common attributes and compared the schemas. Figure 7 summarizes the attributes' presence in the open corporate registers' datasets.

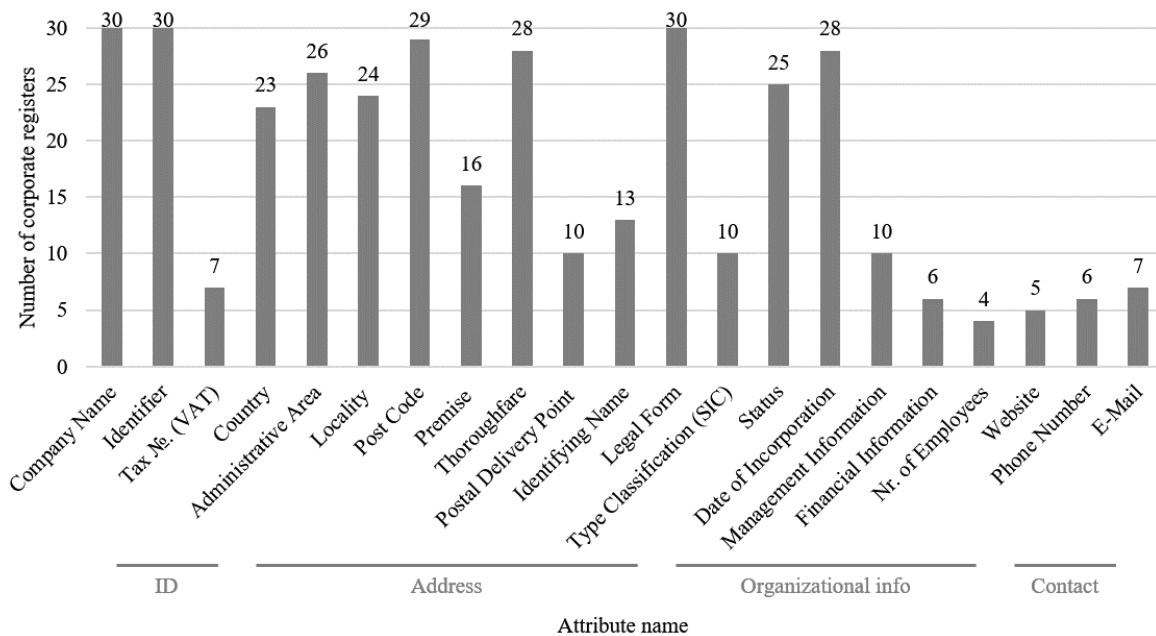


Figure 7. Schema analysis: presence of attributes across the analyzed registers

Companies' address information and their identification concepts are present in the majority of the assessed corporate registers. Nevertheless, certain attributes of these categories, such as "Tax Number", "Premise" and "Postal Delivery Point" appear seldomly. Organizational information

from the business registers in vast majority of the cases provides insights about the incorporation status, date, and companies' legal form, but largely ignores further details about the legal structure and financial statements. Ultimately, such contact detail as "Website", "Phone Number", and "E-mail" are only available for 5 to 7 corporate datasets respectively.

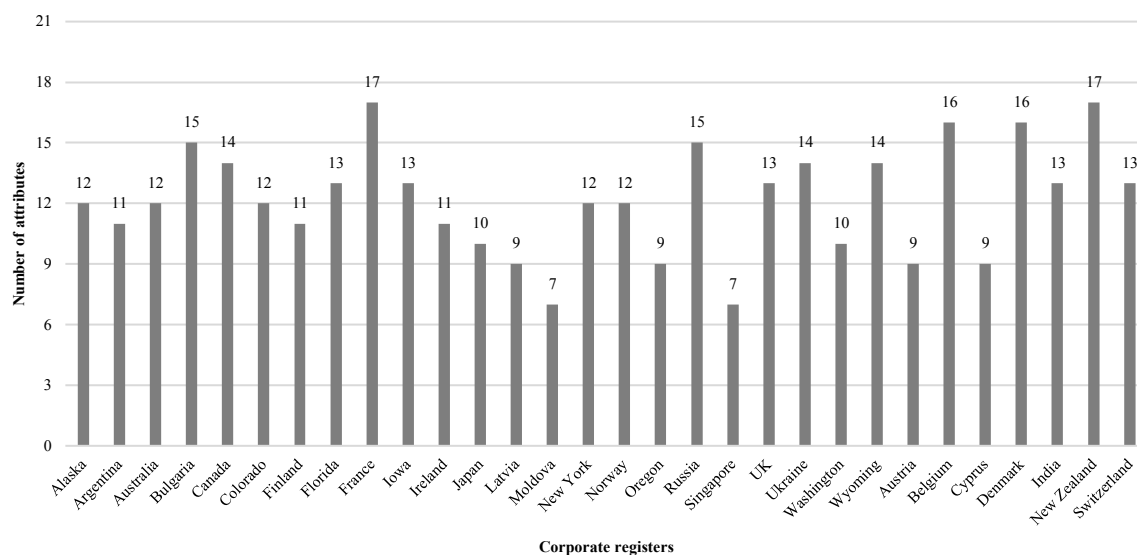


Figure 8. Schema analysis: presence of attributes per corporate register

With a total of 21 attributes, where all of them appear at least once across the datasets, no corporate register is complete. Similar to our previous findings, on average, 12 of the 21 identified attributes were present per register. Figure 8 reveals that the French register accompanied by the New Zealand register have 17 attributes, which is the best result. Belgian and Danish registers show 16 attributes present, followed by Bulgarian and Russian registers. Interestingly the US state registers do not provide the same attributes although they are part of the same country. Business registers of Moldova and Singapore ended up with the least available attributes, yet their data is fully accessible and is published in machine-readable formats (i.e., XLSX and CSV respectively).

4.4 Ready for use assessment

The working sessions described in Section 3 provided practitioners' insights into how OCD can be used in the business environment. These sessions helped us identifying, discussing and validating four concrete use cases, and allowed us to link each attribute to the relevant business concept (see Tables 5-8). For each of the collected use cases, we have marked with an underscore the business concepts which are frequently available in OCD (more than in 80% of the analyzed

corporate registers). This allows us to understand if the data currently available in OCD offers the necessary information for the feasibility of the use cases.

Master Data Management. The maintenance of business partner data (customers and/or vendors) within a company’s IT systems is the most popular OCD use case. This use case supports master data management and aims at maintaining the most accurate version of the data in the company’s internal systems, most prominently for addresses and companies’ names. OCD can help to ensure the data quality by removing duplicates, reconciling concepts representing the same real-world object, enriching the data with new entries, and ensuring its completeness and accuracy by adding up-to-date information from authoritative sources.

This use case, even with an obvious lack of the “Contact” information across corporate registers, is feasible and demonstrates the highest maturity (see Table 3). This is mainly due to the fact that “Identification” and “Address” information are widely present in the registers and are commonly required to be published by the governmental bodies.

Use case benefits	Internal data objects	Attributes from business registers	
		Group	Name
<ul style="list-style-type: none"> • Data quality improvement • Validation of existing records • Duplicate removal • Enrichment with new data • Update and automatic maintenance of data 	<ul style="list-style-type: none"> • Customer master data (ID, address, legal status, contact details) • Vendor master data (ID, address, legal status) • Business partner master data (ID, address, legal status) 	Identification	<u>Company Name</u> , <u>Identifier</u> , Tax Number, (VAT)
		Address	<u>Country</u> , <u>Administrative area</u> , <u>Locality</u> , <u>Post Code</u> , Premise, <u>Thoroughfare</u> , Postal Delivery Point
		Organizational Information	<u>Date of Incorporation</u> , <u>Incorporation Status</u> , <u>Legal Form</u>
		Contact	Website, Phone Number, E-mail

Table 22. Analysis of master data management use cases

Fraud Prevention. OCD can help with the identification of fraudulent business partner. This can be achieved by validating if the business partner counterpart is officially listed in a corporate register. Additionally, in case a register provides details regarding the directors, it can be helpful to verify if the company owner is not present in a sanctions list. OCD can also support investigations into corruption, abuse of power, and violations of cartel laws (Varytimou et al., 2015).

Similar to the previous use case, the presence of “Identification” and “Address” data already allows to identify if the analyzed data entry corresponds to the information provided by the business registers. However, the largely missing “Organizational information” complicates the process of identification of black-listed business partners. The trustworthiness of the data coming from official corporate registers plays a key role for a potential success of this usage scenario, making it a viable use case. Another attribute which would have improved the fraud

identification process are banking details, such as SWIFT or IBAN numbers, related to a particular company. This information was not identified among the assessed corporate registers.

Use case benefits	Internal data objects	Attributes from business registers	
		Group	Name
<ul style="list-style-type: none"> • Reduce fraud risk • Decrease financial losses related to fraud cases • Establish trustworthy relations with business partners 	<ul style="list-style-type: none"> • Business partner master data (ID, VAT, name, address) • Current suppliers and prospects • Banking details, financial structure 	Identification	<u>Company Name</u> , <u>Identifier</u> , Tax Number, (VAT)
		Address	<u>Country</u> , <u>Administrative area</u> , <u>Locality</u> , <u>Post Code</u> , <u>Premise</u> , <u>Thoroughfare</u> , <u>Postal Delivery Point</u> , <u>Identifying name</u>
		Organizational Information	Management Information

Table 23. Analysis of fraud prevention use case

Intelligence and Analytics. OCD can be used to gain insights into customers, partners, and competitors. Moreover, it is possible to identify a particular enterprise with a unique identifier, which helps prevent confusion due to similar company names.

In this regard, companies seek to use more sophisticated information for analytics, which goes beyond the “Address” details. Even though several registers do contain management (33%) and financial information (20%), this is too little to be useful. For instance, only the registers of Denmark and France provide the full set of attributes in this category. We can conclude that this lack of information complicates the feasibility of the Intelligence and Analytics. Consequently, companies willing to pursue this usage scenario are pushed to search for the missing attributes, for instance among specialized data vendors or data marketplaces.

Use case benefits	Internal data objects	Attributes from business registers	
		Group	Name
<ul style="list-style-type: none"> • Improved competitive advantage • Insights about customers, partners, and competitors • Optimization of operational efficiency 	<ul style="list-style-type: none"> • Business partner master data (ID, address, legal status) • Financial structures • Customer contact details 	Identification	<u>Company Name</u> , <u>Identifier</u>
		Organizational Information	<u>Legal Form</u> , <u>Incorporation Status</u> , <u>Date of Incorporation</u> , Number of Employees
		Address	<u>Country</u> , <u>Post Code</u> , <u>Thoroughfare</u> , <u>Identifying Name</u>
		Management Information	Financial Statement, Organizational structure, Number of Employees, <u>Legal Form</u> , Industry Classification Type, <u>Incorporation Status</u>

Table 24. Analysis of intelligence and analytics use case

Marketing. OCD helps to identify potential clients in particular industries and to target marketing campaigns at them. In this case, it is crucial to have up-to-date information about their activities and their initial contact information.

Even though “Company Name” and “Incorporation Status” are among the most available attributes, the rest of the necessary business concepts are far less common. Marketing-related use cases, i.e., marketing campaigns, suffer from a lack of contact information, which is also

relatively scarce in all of the corporate registers. “Contact” category is fully covered in the corporate registers of Bulgaria, Russia, Ukraine, and New Zealand. In this regard, this use case is more difficult to implement in an enterprise setting.

Use case benefits	Internal data objects	Attributes from business registers	
		Group	Name
<ul style="list-style-type: none"> • Reduction of operational costs • Acceleration of procurement activities • Improved analytics 	<ul style="list-style-type: none"> • Business partner master data (ID, address, status, contact details) • Product master data (shipping, tracking, status reports) 	Identification	<u>Company Name</u>
		Organizational Information	<u>Incorporation Status, Industry Classification</u>
		Contact	Website, Postal Delivery Point, Phone Number, E-mail

Table 25. Analysis of marketing use case

It is interesting that “Address” is an overarching concept in all the use cases, while other concepts (identification numbers, organizational information, and contact details) are only relevant for selected use cases. It is widely available in all of the registers, but not all of the attributes are equally present across the datasets. For instance, the complete scope of “Address” information is covered in certain US state registers (Florida, Iowa, New York, and Wyoming) and the business register of Belgium, although important attributes (“Locality”, “Post Code” and “Thoroughfare”) are mostly present. The corporate registers present “Organizational information” only infrequently, with “Contact” details appearing least.

5 Conclusion

Despite governments, NGOs, and companies' enormous efforts to open their data and the open data movement's decade of evolution, the adoption of open data stays generally behind expectations. This is particularly the case for enterprises, which are reluctant to even try open data. Our study contributes to the emerging stream of research on the use of open data and addresses the "lack of insight into the user's perspective" (2012). More specifically, we assess to what extent open corporate data is ready for use in four typical enterprise use cases.

The main contribution of our study is a use case-driven analysis of open corporate registers, which considers both the metadata and the dataset content. Our analysis of 30 corporate registers reveals that open corporate datasets have limited use for typical use cases. On the one hand, the heterogeneity of access, licensing, publishing conditions, and content in open corporate datasets hinder their reuse in a business context. On the other hand, the presence of required business concepts differs per use case. For instance, legally required information about companies, such as their addresses and identification, is mostly available, but not always complete, while many other attributes are only partially available. Therefore, the most interesting insights from our study are the ready for use assessments for four specific use cases. We find that master data management can already benefit from OCD, whereas the other use cases lack the required business concepts. To the best of our knowledge, the conducted analysis is one of the first to provide insights into open data's readiness for use from an enterprise perspective.

Beyond the assessment of OCD, our study provides a methodological contribution by proposing a use case-driven approach comprising four steps: (1) the identification of the open data sources, (2) a metadata analysis, (3) a schema analysis of the datasets, and (4) a ready for use assessment based on a comparison to relevant business concepts in the selected use case. This approach goes beyond the existing assessment approaches of open data quality (see 2.2) by integrating the use context and business scenario.

A limitation of this work is that our analysis focuses on selected registers in countries that are considered as advanced with regard to open data provision (Publications Office of the EU, 2020). Given the total number of existing business registers, our sample does not allow us to draw conclusions about the domain as a whole. In addition, our assessment relies on four use cases identified by the focus group, but others could be potentially discovered.

An implication from our study is that the proposed open data assessment methods require amendments to integrate the user's perspective. Future research needs to put more emphasis on domain- and use case-specific analysis to complement these methods in order to assess open data's usability. We also see opportunities to further develop the proposed approach to cover other open data domains. This could result in developing a general approach to usability assessment for open data from the enterprise perspective. From a theoretical perspective the concept of open data quality should be revisited with regard to usability (Bicevskis et al., 2018; Vetrò et al., 2016). In order to thoroughly address the data quality aspects, future research could embed the assessment techniques with metrics along the data quality dimensions in the content analysis part (e.g., timeliness, accuracy, and completeness).

Our study also underlines the need for domain ontologies, such as the euBusinessGraph (2019) common semantic model for company data, which could be a basis to provide more consistent and compatible open datasets across different open data portals and providers.

6 References

- Barry, E., & Bannister, F. (2014). Barriers to Open Data Release: A View from the Top. *Information Polity*, 19(1/2), 129–152.
- Beno, M., Figl, K., Umbrich, J., & Polleres, A. (2017). Open Data Hopes and Fears: Determining the Barriers of Open Data. *Proceedings of the 2017 Conference for E-Democracy and Open Government*, 69–81.
- Bicevskis, J., Bicevska, Z., Nikiforova, A., & Oditis, I. (2018). Data Quality Evaluation: A Comparative Analysis of Company Registers' Open Data in Four European Countries. *Communication Papers of the 2018 Federated Conference on Computer Science and Information Systems*, 17, 197–204.
- Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data—The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3), 1–22.
- Bogdanović-Dinić, S., Veljković, N., & Stoimenov, L. (2014). How Open Are Public Government Data? An Assessment of Seven Open Data Portals. In *Measuring E-government Efficiency* (Vol. 5, pp. 25–44). Springer.
- Braunschweig, K., Eberius, J., Thiele, M., & Lehner, W. (2012). The State of Open Data Limits of Current Open Data Platforms. *Proceedings of 21st World Wide Web Conference*.
- Bryman, A., & Bell, E. (2007). *Business Research Methods* (2nd ed.). Oxford University Press.
- Conradie, P., & Choenni, S. (2014). On the barriers for local government releasing open data. *Government Information Quarterly*, 31, S10–S17.
- Creswell, J. W. (2009). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (3rd ed.). SAGE Publications.
- Davies, T., Walker, S., Rubenstien, M., & Perini, F. (Eds.). (2019). *The State of Open Data: Histories and Horizons*. African Minds and International Development Research Centre. <https://www.idrc.ca/en/book/state-open-data-histories-and-horizons>
- Deloitte Analytics. (2012). *Open Data – Driving Growth, Ingenuity and Innovation*. <https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/deloitte-analytics/open-data-driving-growth-ingenuity-and-innovation.pdf>
- Dinter, B., & Kollwitz, C. (2016). Towards a Framework for Open Data Related Innovation Contests. *Proceedings of the 2016 Pre-ICIS SIGDSA/IFIP WG8.3 Symposium: Innovations in Data Analytics*, 13.
- euBusinessGraph. (2019). Ontology for Company Data. *EuBusinessGraph*. <https://www.eubusinessgraph.eu/eubusinessgraph-ontology-for-company-data/>
- European Commission. (2020). *Open Data Maturity Report 2019*. https://op.europa.eu/publication/manifestation_identifier/PUB_OABE19001ENN
- European Commission, Capgemini Consulting, Intrasoft International, Fraunhofer Fokus, con.terra, Sogeti, Open Data Institute, Time.lex, & University of Southampton. (2015). *Creating Value through Open Data Study on the Impact of Re-use of Public Data Resources*. Publications Office of the European Union.
- European Parliament. (2012). *Directive 2012/17/EU* [Text]. Queen's Printer of Acts of Parliament. <https://www.legislation.gov.uk/eudr/2012/17/body>
- GLEIF. (2017). *Accreditation Process*. GLEIF. <https://www.gleif.org/en/about-lei/gleif-accreditation-of-lei-issuers/accreditation-process>
- GLEIF. (2019). *GLEIF Registration Authorities List*. GLEIF. <https://www.gleif.org/en/about-lei/code-lists/gleif-registration-authorities-list>
- Global Open Data Index. (2015). *Company Register*. <https://index.okfn.org/dataset/companies/>
- Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, 29(4), 258–268.
- Kampars, J., Zdravkovic, J., Stirna, J., & Grabis, J. (2020). Extending Organizational Capabilities with Open Data to Support Sustainable and Dynamic Business Ecosystems. *Software and Systems Modeling*, 19(2), 371–398.
- Koznov, D., Andreeva, O., Nikula, U., Maglyas, A., Muromtsev, D., & Radchenko, I. (2016). A Survey of Open Government Data in Russian Federation: *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 173–180.
- Krasikov, P., Harbich, M., Legner, C., & Eurich, M. (2019). *Open Data Use Cases: Framework for the Generation and Documentation of Open Data Use Cases* [CC CDQ working report]. https://www.cc-cdq.ch/system/files/Open_data_use_cases_working_report_2019.pdf
- Krasikov, P., Obrecht, T., Legner, C., & Eurich, M. (2020). Is Open Data Ready for Use by Enterprises? *Proceedings of the 9th International Conference on Data Science, Technology and Applications*, 109–120.
- Kubler, S., Robert, J., Neumaier, S., Umbrich, J., & Le Traon, Y. (2018). Comparison of Metadata Quality in Open Data Portals Using the Analytic Hierarchy Process. *Government Information Quarterly*, 35(1), 13–29.
- Máchová, R., & Lněnička, M. (2017). Evaluating the Quality of Open Data Portals on the National Level. *Journal of Theoretical and Applied Electronic Commerce Research*, 12, 21–41.
- Manyika, J., Chui, M., Groves, P., Farrell, D., Van Kuiken, S., & Doshi, E. A. (2013). *Open Data: Unlocking Innovation and Performance with Liquid Information*. McKinsey Global Institute. <https://www.mckinsey.com/business->

- functions/digital-mckinsey/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information
- Martin, S., Foulonneau, M., Turki, S., & Ihadjadene, M. (2013). *Risk Analysis to Overcome Barriers to Open Data*. 11(1), 348–359.
- Neumaier, S., Umbrich, J., & Polleres, A. (2016). Automated Quality Assessment of Metadata across Open Data Portals. *Journal of Data and Information Quality*, 8(1), 1–29.
- Oliveira, M. I. S., Oliveira, L. E. R. de A., Lima, G. de F. A. B., & Lóscio, B. F. (2016). Enabling a Unified View of Open Data Catalogs: *Proceedings of the 18th International Conference on Enterprise Information Systems*, 230–239.
- Open Government Working Group. (2007). *The 8 Principles of Open Government Data*. <https://opengovdata.org/>
- OpenCorporates. (2020). *Open Company Data Index*. <http://registries.opencorporates.com/>
- OpenDataBarometer & World Wide Web Foundation. (2020, October). *Open Data Barometer*. https://opendatabarometer.org/?_year=2017&indicator=ODB
- Osagie, E., Waqar, M., Adebayo, S., Stasiewicz, A., Porwol, L., & Ojo, A. (2017). Usability Evaluation of an Open Data Platform. *Proceedings of the 18th Annual International Conference on Digital Government Research*, 495–504.
- Puha, A., Rinciog, O., & Posea, V. (2018). Enhancing Open Data Knowledge by Extracting Tabular Data from Text Images: *Proceedings of the 7th International Conference on Data Science, Technology and Applications*, 220–228.
- Reiche, K. J., Höfig, E., & Schieferdecker, I. (2014). Assessment and Visualization of Metadata Quality for Open Government Data. *Conference for E-Democracy and Open Governement*, 335–346.
- Stróżyńska, M., Eiden, G., Abramowicz, W., Filipiak, D., Małyszko, J., & Węcel, K. (2018). A Framework for the Quality-based Selection and Retrieval of Open Data. *Electronic Markets*, 28(2), 219–233.
- Umbrich, J., Neumaier, S., & Polleres, A. (2015). Quality Assessment and Evolution of Open Data Portals. *2015 3rd International Conference on Future Internet of Things and Cloud*, 404–411.
- Varytimou, A., Loutas, N., & Peristeras, V. (2015). Towards Linked Open Business Registers: The Application of the Registered Organization Vocabulary in Greece. *International Journal on Semantic Web and Information Systems*, 11(2), 66–92.
- Vetrò, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., & Morando, F. (2016). Open Data Quality Measurement Framework. *Government Information Quarterly*, 33(2), 325–337.
- Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, 12(4), 5–33.
- Weerakkody, V., Irani, Z., Kapoor, K., Sivarajah, U., & Dwivedi, Y. K. (2017). Open data and its usability: An empirical view from the Citizen's perspective. *Information Systems Frontiers*, 19(2), 285–300.
- Welle Donker, F., & Van Loenen, B. (2017). How to Assess the Success of the Open Data Ecosystem? *International Journal of Digital Earth*, 10(3), 284–306.
- Wikipedia. (2019). *List of company registers*. Wikipedia. https://en.wikipedia.org/w/index.php?title=List_of_company_registers&oldid=922064632
- Zhang, R., Indulska, M., & Sadiq, S. (2019). Discovering Data Quality Problems: The Case of Repurposed Data. *Business & Information Systems Engineering*, 61(5), 575–593.
- Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R., & Alibaks, R. S. (2012). Socio-Technical Impediments of Open Data. *Electronic Journal of E-Government*, 10(2), 156–172.

Appendix 1

Source Information		
ID	Registry Code	Unique identification of legal entities by GLEIF (2019).
	Country	Defines a country to which the register refers.
Access	Resource Format	Describes the format of the published data, e.g., XML, JSON, CSV.
	Access Login	Mentions whether access to the dataset requires an account.
	Free Lookup Service	Indicates whether the register has a free company lookup service.
	License	License under which the data is provisioned.
Publisher	Publisher	Entity responsible for providing the data.
	Publishing Date	Date when the register originally published the dataset.
	Update Cycle	Describes the frequency of the data update in days.
Content	Resource Language	Mentions the language(s) in which the data is published.
	Geographic Coverage	Defines the scope of the publishing institution, either on a state or national level.
	# of Diverse Attributes	Counts the different attributes that the register reports.
	#of Records	Estimate of the total number of entries in a register.
Content Information		
ID	Company Name	Defines the entity's name in a local language.
	Identifier	A unique identifier assigned to the relevant register.
	Tax N° (VAT)	The tax number of the entity (VAT).
Address	Country	A geopolitical area, typically a nation.
	Administrative Area	A top-level geographical or political area division in a country.
	Locality	A more granular level of an administrative area's geographical division.
	Post Code	A country-specific code for a certain address component.
	Premise	An area of land and its adjacent buildings.
	Thoroughfare	A form of the access route of the address: a street, road, avenue, etc.
	Postal Delivery Point	A single mailbox or other place at which postal mail is delivered.
	Identifying Name	A name assigned to an address, e.g., the legal representative.
	Legal Form	The type of entity with respect to the local corporate law.
	Type Classification (SIC)	Classification of entities and their respective industries.
Organizational information	Status	The entity's status, e.g., active, bankrupt.
	Date of Incorporation	Date of the entry in the register.
	Management Information	Information about the company's organizational structure and legal ownership.
	Financial Information	Usually financial reports on corporate figures.
	Number of Employees	The entity's number of employees.
Contact	Website	The entity's website.
	Phone Number	The entity's phone number.
	E-Mail	The e-mail address of the entity.

Table 26. Definition of attributes

Appendix 2

		Alaska 1	Argentina 2	Australia 3	Bulgaria 4	Canada 5	Colorado 6	Finland 7	Florida 8	France 9	Iowa 10
ID	Registry Code	RA000594	RA000010	RA000013	RA000065	RA000072	RA000599	RA000188	RA000603	RA000189	RA000606
	Country	United States	Argentina	Australia	Bulgaria	Canada	United States	Finland	United States	France	United States
Access	Resource Format	CSV	CSV	XML, SOAP API	XML, JSON	XML, API	CSV, RDF, RSS, TSV, XML, REST	JSON, API, HTTP	TXT	CSV, API	CSV, RDF, RSS, TSV, XML, SODA API
	Access Login Free	no	no	no	no	no	no	no	no	no	no
	Lookup Service	available	available	available	available	available	available	available	available	not available	available
License	License	Open Government License	Creative Commons Attribution 4.0	Creative Commons Attribution 3.0 Australia	Open License	Open Government License - Canada	Public Domain	Creative Commons Attribution 4.0	N/A	Open License V2.0	Creative Commons Attribution 4.0
Publisher	Publisher	State of Alaska Department of commerce	Argentinian Ministry of Justice and Human Rights	Australian Business Register	Bulgarian Ministry of Justice Registry Agency	Innovation, Science and Development Canada	Colorado Department of State	Finnish Patent and Registration Office	Division of Corporation Florida	National Institute of Statistics and Economic Studies	Secretary of State Iowa
	Publishing Date	N/A	19 Sep 2016	05 Sep 2014	N/A	18 Feb 2014	19 Mar 2014	N/A	N/A	24 Aug 2018	10 Nov 2014
	Update Cycle	N/A	30d	1d	N/A	7d	1d	N/A	1d	1d	30d
Content	Resource Language	English	Spanish	English	Bulgarian	English, French	English	English, Finnish	English	French, English	English
	Geographic Coverage	State	National	National	National	National	State	National	State	National	State
	# of Records	82,902	N/A	18,000,000	972,362	995,900	2,043,641	874,382	8,948,976	21,059,740	260,522
# of Diverse Attributes	35	26	22	11	25	35	86	45	118	19	

Figure 9. Metadata analysis of corporate registers 1 to 10

		Ireland 11	Japan 12	Latvia 13	Moldova 14	New York 15	Norway 16	Oregon 17	Russia 18	Singapore 19	UK 20
ID	Registry Code	RA000402	RA000413	RA000423	RA000451	RA000628	RA000472	RA000631	RA000499	RA000523	RA000585
	Country	Ireland	Japan	Latvia	Moldova	United States	Norway	United States	Russia	Argentina	United Kingdom
Access	Resource Format	REST API	XML, CSV (Shift_JIS), CSV (Unicode)	CSV, XLSX	XLSX	CSV, RDF, RSS, TSV, XML	CSV, JSON, XML, REST API	CSV, RDF, RSS, JSON, XML, SODA API	XML	CSV	CSV, REST
	Access Login Free	no	no	no	no	no	no	no	no	no	no
	Lookup Service	available	available	available	available	available	available	available	available	available	available
License	License	Open License	Open License	N/A	N/A	Open Government License	Norwegian Open License	N/A	Open License	Singapore Open Data License	Free, Open Government License v3.0
Publisher	Publisher	Companies Registration Office Ireland	Financial Service Agency	Register of Enterprises of the Republic of Latvia	Moldavian State Chamber of Registration	New York Department of State	The Central Coordinating Register for Legal Entities	Secretary of State Oregon	Russian Federal Tax Service	Singapore Accounting and Corporate Regulatory Authority	Companies House UK
	Publishing Date	N/A	N/A	10 Mar 2014	N/A	14 Feb 2013	N/A	19 May 2016	01 Aug 2016	12 Dec 2016	11 Dec 2013
	Update Cycle	N/A	30d	N/A	30d	30d	N/A	7d	7d	Ad-hoc	7d
Content	Resource Language	English	Japanese, English	Latvian	Romanian	English	Norwegian	English	Russian	English	English
	Geographic Coverage	National	National	National	National	State	National	State	National	National	National
	# of Records	N/A	4,937,210	425,637	223,841	2,944,438	1,823,057	397,816	N/A	1,613,261	12,649,839
# of Diverse Attributes	18	19	21	10	30	42	18	115	8	55	

Figure 10. Metadata analysis of corporate registers 11 to 20

		Ukraine 21	Washington 22	Wyoming 23	Austria 24	Belgium 25	Cyprus 26	Denmark 27	India 28	New Zealand 29	Switzerland 30
ID	Registry Code	RA000567	RA000641	RA000644	RA000687	RA000025	RA000161	RA000170	RA000394	RA000466	RA000548
	Country	Ukraine	United States	United States	Austria	Belgium	Cyprus	Denmark	India	New Zealand	Switzerland
Access	Resource Format	XML	XML, JSON, TXT	CSV	PDF	PDF	WebGUI	REST API	CSV	XML, JSON, API	WebGUI
	Access Login	no	no	no	yes	yes	yes	yes	yes	yes	no
	Free Lookup Service	not available	available	available	available	available	available	available	not available	available	available
License	License	Open License	N/A	N/A	restricted access	restricted to queries	Open License	N/A	National License	Creative Commons Attribution 4.0	restricted to queries
Publisher	Publisher	Ukrainian National Information Systems	Secretary of State Washington	Secretary of State Wyoming	Federal Ministry Republic of Austria Digital and Economic Affairs	Ministry of Economy Belgium	Cyprus Department of Registrar of Companies and Official Receiver	Danish Business Authority	Ministry of Corporate Affairs India	Ministry of Business Innovation and Employment of New Zealand	Swiss Federal Statistical Office
	Publishing Date	12 Dec 2016	N/A	19 Mar 2014	N/A	N/A	N/A	10 Jun 2015	N/A	N/A	11 Dec 2013
	Update Cycle	5d	1d	N/A	N/A	7d	N/A	1d	365d	N/A	1d
Content	Resource Language	Ukrainian	English	English	German, English	English, French, Dutch, German	English, Greek, Turkish	Danish, English, Kalaallisut	English	English	English, French, Italian, German
	Geographic Coverage	National	State	State	National	National	National	National	State	National	National
	# of Records	1,743,903	1,381,897	522,691	N/A	1,235,529	425,060	N/A	N/A	N/A	N/A
	# of Diverse Attributes	131	20	87	14	22	11	35	17	50	25

Figure 11. Metadata analysis of corporate registers 21 to 30

Essay 4

A Method to Screen, Assess, and Prepare Open Data for Use

Pavel Krasikov and Christine Legner

Faculty of Business and Economics (HEC), University of Lausanne, Switzerland

*First version published in the Proceedings of the 16th International Conference on Design Science
Research in Information Systems (DESRIST), 2021*

*Extended version accepted to the special issue of the Journal of Data and Information Quality
(JDIQ) on Quality Aspects of Data Preparation, 2023*

Abstract: *Open data's value-creating capabilities and innovation potential are widely recognized, resulting in a notable increase in the number of published open data sources. A crucial challenge for companies intending to leverage open data is to identify suitable open datasets that support specific business scenarios and prepare these datasets for use. Researchers have developed several open data assessment techniques, but those are restricted in scope, do not consider the use context, and are not embedded in the complete set of activities required for open data consumption in enterprises. Therefore, our research aims to develop prescriptive knowledge in the form of a meaningful method to screen, assess, and prepare open data for use in an enterprise setting. Our findings complement existing open data assessment techniques by providing methodological guidance to prepare open data of uncertain quality for use in a value-adding and demand-oriented manner, enabled by knowledge graphs and linked data concepts. From an academic perspective, our research conceptualizes open data preparation as a purposeful and value-creating process.*

Keywords: Open data, Data preparation, Data quality, Action design research, Knowledge graph

Table of contents

1	Introduction.....	131
2	Prior research.....	133
2.1	Open data and adoption barriers.....	133
2.2	Open data quality and assessment techniques.....	134
2.3	Open data processes form publisher and consumer perspectives.....	136
2.4	Research gap.....	138
3	Methodology.....	139
3.1	Research objectives and setting.....	139
3.2	Research process.....	139
4	Method to screen, assess, and prepare open data for use.....	142
4.1	Purpose and design considerations.....	142
4.2	Phases and illustration.....	143
4.3	Workflow.....	154
5	Comparison with other frameworks and approaches.....	157
6	Conclusion and limitations.....	158
7	References.....	160

List of figures

Figure 12. Example of the documentation of an open dataset (from https://meta.cdq.com/Data_source/FR.RC).....	154
Figure 13. Example of reference ontology and entity-linking process for selected datasets.....	154
Figure 14. Possible variations in the workflow of the method to screen, assess, and prepare open data for use.....	156

List of tables

Table 27. Main barriers to open data adoption in enterprises and their impact on open data consumption.....	134
Table 28. Open data assessment techniques	135
Table 29. Publishers' and consumers' perspectives on open data processes.....	137
Table 30. BIE cycles and their contribution to method development	141
Table 31. Example of use case ideation	144
Table 32. Overview of the method to screen, assess, and prepare open data for use	145
Table 33. Example of identified open data sources and datasets.....	147
Table 34. Relevant open data quality dimensions on metadata, schema and dataset content level, based on Neumaier et al. (2016), Vetrò et al. (2016), and Zhang et al. (2019).	148
Table 35. Examples of the datasets' assessment results on the metadata, schema, and content levels	152
Table 36. Comparison with other approaches	157

1 Introduction

Open data is known to be free for use, reuse, and redistribution by anyone (Open Knowledge Foundation, 2005). It offers business and innovation potential to companies and national economies (Janssen et al., 2012; Zuiderwijk et al., 2015), with an estimated total market size in the European Union of 325 billion euros (European Commission et al., 2015). As the availability of open data sources increases, so do companies' expectations toward open data to fuel advanced analytics, optimize business processes, enrich data management, or even enable new services (Zuiderwijk et al., 2015; Schatsky et al., 2019; Enders et al., 2021). However, as simple and effortless as the free availability of open data may appear, open data consumers have to overcome significant hurdles to identify suitable datasets and prepare them for use in the enterprise context. These barriers hinder companies from leveraging open data's value generating potential (Enders et al., 2020) and lead to a "mismatch between the needs and expectations of the users and the possibilities offered by available datasets" (Ruijter et al., 2018), with the result that the actual use of open data falls short of expectations.

Many of these hurdles are associated with data quality issues, e.g., a lack of transparency about a dataset's content, incomplete or missing data, or unclear licensing and access conditions (Bachtar et al., 2020; Krasikov et al., 2020; Vetrò et al., 2016). To address these issues, researchers have developed dedicated assessment techniques, such as the "Luzzu" framework (Debattista et al., 2016), the "LANG" approach (R. Zhang et al., 2019), or the "QUIN" usability criteria (Osagie et al., 2017). However, these techniques are limited in their assessment scope and mostly consider only the metadata level. Moreover, these techniques are not embedded in the complete set of activities required for open data consumption in enterprises. For instance, they are poorly linked to data preparation, which includes techniques such as data collection, data integration, data transformation, and data cleaning (S. Zhang et al., 2003). To the best of our knowledge, suitable processes and methodological approaches that help prepare open data for enterprise use do not yet exist, at least not in a well-structured, holistic, and rigorous scientific manner. It therefore remains uncertain which process steps and actions qualify to identify, assess, and prepare open data for use successfully.

For this reason, our study focuses on the enterprise setting of open data use, which has not been explicitly addressed in previous studies, and on open data's context-aware quality assessment and preparation, as a prerequisite for the productive use of open data. This leads to the research question:

How can companies be helped to systematically screen, assess, and prepare open data for use?

In line with the principles of Action Design Research (Sein et al., 2011), we engaged with enterprises to understand their current issues and requirements regarding open data use and iteratively developed a method to address them. Our proposed method ensures a purposeful discovery and selection of open data sources and datasets, with consideration of relevant aspects such as provenance, licensing, and access conditions. It integrates a systematic approach to quality assessment of open datasets, being a major criterion for their selection and preparation for further use. This article presents an extended and revised version of an earlier version of the method (Krasikov, Legner, et al., 2021) that was published in the Proceedings of the 16th International Conference of Design Science Research in Information Systems and Technology (DESRIST 2021). Compared to the previous version, we refine and extend the formulated method and its phases, paying particular attention to open data assessment as an essential part of preparation for use.

For the scientific community, our method enriches the existing body of knowledge on open data assessment (see subsection 2.2), by suggesting a three-step approach to context-aware quality assessment. The method also contributes to literature on open data processes (see subsection 2.3) by outlining four process phases and the underlying techniques that qualify to identify, assess, and prepare open data for use successfully. In addition, the proposed method facilitates the systematic analysis and integration of open datasets, thereby conceptualizing open data preparation as a meaningful value-creating process. The method can also serve as a framework for future research; academics can use it to allocate research activities along its various phases or to instantiate it for specific open data use cases.

The remainder of this paper is structured as follows: Section 2 introduces the related work. Section 3 elaborates on our research objectives and the research process. Section 4 presents our method to screen, assess, and prepare open data for use, followed by section 5 which compares it with existing frameworks and approaches. In section 6, we summarize and discuss our findings and present the limitations and outlook on future work.

2 Prior research

Open data is most often associated with but not limited to open government data. Numerous national open data initiatives have produced almost 4000 available open data portals worldwide (Opendatasoft, 2022), with data.europa.eu and data.gov combined providing access to more than 1.7 million open datasets (Data.gov, 2022; EU Open Data Portal, 2022). Despite these impressive numbers, open data use by enterprises remains below expectations (Zuiderwijk et al., 2015). Prior research has investigated barriers to open data adoption – data quality being among the most widespread (subsection 2.1), developed dedicated techniques for open data quality assessment (subsection 2.2), and proposed open data publishing and consumption processes (subsection 2.3).

2.1 Open data and adoption barriers

Contrary to the widespread perception that open data only comprises public information assets published by official authorities, it actually refers to any type of data that is “freely available and can be used as well as republished by everyone without restrictions from copyright or patents” (Braunschweig et al., 2012). One of the major misconceptions about open data (Janssen et al., 2012) is the assumption that simply providing access to data is sufficient for its successful reuse. Open data platforms and their features are known as facilitators to open data use (Zuiderwijk et al., 2012; Bizer et al., 2009; Auer et al., 2007; Zaveri et al., 2016; R. Zhang et al., 2019), but they remain insufficient and have been criticized in terms of functionalities, namely in the public sector (Corsar & Edwards, 2017; Marmier & Mettler, 2020). Although existing open data literature has identified a large set of barriers (Janssen et al., 2012), three main categories stand out as barriers for the enterprise use of open data (Krasikov et al., 2020): a lack of transparency, heterogeneity, and the unknown quality of open datasets. The first barrier (transparency) refers to the difficulties of identifying “the right data” (Janssen et al., 2012), as well as to the understanding of its content and the consistency of conclusions drawn when analyzing it. The second barrier (heterogeneity) challenges the discrepancies of how open data is made available in terms of file formats, data structure, as well as access conditions, licenses, and use permissions (Martin et al., 2013; Zuiderwijk et al., 2012). The third barrier (quality) mentions the deficient information quality of open datasets on multiple levels: inaccurate or incomplete data and obsolete or non-valid records (Janssen et al., 2012; Krasikov et al., 2020). Table 27 synthesizes the main categories of barriers and their impact on enterprises as open data consumers.

Category	Description	Impact on enterprises	Sources
Transparency	Unclear content of the data with a lack of transparency concerning the content, mainly driven by publishers' reluctance to provide clear descriptions of and information about the provided data.	Difficulties in identifying "the right data" and understanding the content and possible use contexts.	Janssen et al., (2012); Zuiderwijk et al., (2012)
Heterogeneity	Variety of forms in which open data is made available, particularly heterogenous structures and formats.	Significant efforts for harmonization of file formats, and data structures. Uncertainty about licensing and use permissions.	Janssen et al., (2012); Zuiderwijk et al., (2012); Martin et al., (2013); Conradie and Choenni, (2014); Barry and Bannister, (2014)
Quality	Unclear quality of the data, i.e., essential information is missing or incomplete, obsolete or non-valid data, and similar data made available by different publishers but yielding different results when analyzed.	Lack of trust in open data as well as limited usefulness and use. Significant efforts for data quality assessment and data preparation.	Janssen et al., (2012); Zuiderwijk et al., (2012); Conradie and Choenni, (2014); Beno et al., (2017); Corsar and Edwards, (2017)

Table 27. Main barriers to open data adoption in enterprises and their impact on open data consumption

2.2 Open data quality and assessment techniques

To overcome the quality-related barriers, researchers have developed dedicated assessment techniques that aim to provide quality metrics and identify data quality issues of open data. While the open data assessment literature is quite extensive (see Table 28), the suggested techniques differ in the scope of the assessment and the methodologies used by the authors.

Regarding assessment scope, it is evident that the assessment of metadata's quality at the source level is the center of attention. A main reason for the focus on metadata is the discoverability of open datasets, which purport the importance of understanding the open data's content before using it. The few papers that focus their assessment scope on datasets (Debattista et al., 2016; Vetrò et al., 2016; R. Zhang et al., 2019) are inspired by classical methodologies on data quality assessment, especially those proposed by Batini et al. (2009) and Pipino et al. (2002). Interestingly, these papers propose universal approaches that are formulated independently of the use context, whereas seminal data quality literature emphasizes the subjective use-oriented view of quality (Corsar & Edwards, 2017). Hence, although the open data assessment literature provides a clear link to the traditional data quality literature (R. Zhang et al., 2019), it neglects the open data consumers' perspective (Krasikov, Obrecht, et al., 2021). We argue that the definition of data quality, commonly referred to as "fitness for use" (Richard Y. Wang & Diane M. Strong, 1996), must equally apply to open data, emphasizing the importance of open data's "usefulness" in specific use cases (Osagie et al., 2017), and not only its usability from a technical

standpoint. To this end, traditional data quality metrics play an essential role in preparing open data for further use, but their sufficiency and context considerations remain unaddressed.

Source	Assessment approach	Assessment scope	Methodology
Bogdanović-Dinić et al., (2014)	“Data openness score” based on eight open data principles (Open Government Working Group, 2007)	Metadata	Case study: application of the “data openness” model to 7 open data portals
Reiche et al., (2014)	Ranking of open data repositories according to the average score computed by means of quality metrics	Metadata	Case study: assessment of the metadata quality of 10 open government data portals
Debattista et al., (2016)	Framework “Luzzu”, to assess linked open data quality along the 22 dimensions based on RDF vocabularies	Metadata and dataset	Literature-based definition of the quality metrics for the methodology; evaluation performed on 9 datasets from “270a” data space
Neumaier et al., (2016)	Metadata quality assessment framework with 29 dimensions derived from DCAT	Metadata	Assessment of 261 open data portals to highlight common issues
Vetrò et al., 2016 (2016)	Quality framework supported by data quality models from the literature, with 6 dimensions and 14 metrics	Metadata and dataset	Quantitative assessment of the quality of 11 datasets, supported by data quality models from the literature
Máchová and Lněnička, (2017)	Benchmarking framework to evaluate open data portals’ quality, with 12 general characteristics and 16 metrics	Metadata	Quality evaluation of 67 open data portals
Welle Donker and van Loenen, (2017)	Holistic open data assessment framework with 3 main levels: open data supply, open data governance, and open data user characteristics	Metadata	Assessment of 20 “most wanted” datasets addressing open data in the Netherlands
Osagie et al., 2017 (2017)	Usability evaluation “QUIN” criteria (12 usability criteria)	Platform features	Evaluation as part of the agile development process “ROUTE-TO-PA”
Bicevskis et al., (2018)	Three-part data quality model: definition of a data object, data object quality specifications, and implementation	Dataset	Syntax analysis of data from 4 company registers for 11 attributes
Stróżyńska et al., (2018)	Quality-based selection, assessment, and retrieval method	Metadata	Attribution of quality scores based on “ranking type Delphi” and 6 quality dimensions to 59 data sources
Zhang et al., 2019 (2019)	Discovery of data quality problems in 20 datasets using the “LANG” approach, according to 10 dimensions	Metadata and dataset	Design science research and a systematic approach to repurposed datasets’ quality
Nayak, Bozic, and Longo, (2021)	Ontological approach to report data quality violated triples, including an assessment and root cause analysis with 17 metrics	Metadata	Qualitative study on linked open data assessment, based on the existing literature

Table 28. Open data assessment techniques

2.3 Open data processes from publisher and consumer perspectives

While open data quality assessment techniques focus on metadata and the data itself, another research stream addresses the processes associated with the publishing and use of open data. These studies predominantly target open data publishers and focus on the identification and selection processes of the data to be published (see Table 29). Only two of the existing studies address the processes exclusively from the consumers' perspective (Hendler, 2014; Zuiderwijk et al., 2015). Even though the contexts of these papers differ, they outline similar processes for open data users, namely finding (identifying), analyzing, and processing (integrating and validating) open data.

Ren and Glissmann (2012) propose a five-phase process to identify open data information assets to drive open data initiatives. This structured approach, adopting a governmental perspective, focuses on concrete steps to harvest value from open data: define business goals, identify stakeholders, identify potential information assets, assess quality, and select information assets. Although this approach does not reflect a user perspective, the authors regard the selection of information assets as a key decision that ensures the subsequent positive impact of open data use. They also highlight the need for guidelines that could increase publishers' return on investment when engaging in open data initiatives.

Zuiderwijk and Janssen (2014) investigate sociotechnical barriers and developments in open data processes from both perspectives – publishers (governments) and users (citizens) – along with six highly dependent steps for the open data processes: creating, opening, finding, analyzing, processing, and discussing. While creating and publishing open data refer to data providers, open data consumers are involved in the finding and using steps. The authors conclude: “the data that are published are usually not published in a format that makes it easy to reuse the data” (Zuiderwijk et al., 2014).

Source	Perspective and context	Research method	Processes (publishers)	Processes (consumers)
Ren and Glissmann, (2012)	Open data publisher (government) Identifying and incorporating information assets for open data initiatives	Based on principles of business architecture and information quality	Define business goals, identify stakeholders, identify potential information assets, assess readiness, and select information assets	N/A

Source	Perspective and context	Research method	Processes (publishers)	Processes (consumers)
Masip-Bruin et al, (2013)	Open data publisher (city council) Systematic value creation process, enabled by middleware, to identify suitable information to be used	Scenario and practice driven	Data selection, acquisition, and processing	N/A
Zuiderwijk and Janssen, (2014)	Open data publisher (government) and user Sociotechnical impediments of open data along the high-level representation of open data processes	Literature review (n=37), semi-structured interviews (n=6), workshops (n=4), and a questionnaire (300 respondents)	Governmental organizations: create, open, and publish data	Users: find, analyze, and process open data
			Both: discuss and provide feedback	
Hendler, (2014)	Big data user Integration techniques for structured and unstructured online data, exemplified with open data	Explorative analysis	N/A	Discover, integrate, and validate open datasets
Zuiderwijk et al., (2015)	Open data user Commercial open data use to create a competitive advantage	Multi-method study: scenario development, semi-structured interviews (n=2), and a survey (n=14).	N/A	Search for open data, find open data, use open data, enrich open data, and link it to internal datasets, interpret findings, and draw conclusions
Crusoe and Melin, 2018 (2018)	Open data publisher (government) and user Investigating and systematizing open government data research	Literature review (n=34)	Governmental organizations: identify data suitability, take release decisions, publish open data, evaluate the impact, and collect feedback	End users: use open data and provide feedback
Abella et al., (2019)	Open data publisher and user Impact generation process of open data	Practice-driven analysis	Organizations: qualify data for publication, publish open data	External: reuse open data
			Open data reuse generates impact	
Abida et al., (2020)	Open data publisher Integrating and publishing linked open government data	Illustrative case study	Data transformation, interlinking, storage, visualization, and publishing	N/A

Table 29. Publishers' and consumers' perspectives on open data processes

Continuing the exploration of open data barriers, Crusoe and Melin (2018) expand the open government data process (Zuiderwijk et al., 2014), where publishers are additionally involved in assessing the suitability of open data, and releasing it. From the users' perspective, open datasets lack contextual interpretations, are difficult to find, are hard to understand, and often do not consider the needs of open data users. Businesses are often positioned as both publishers and

consumers of open data (Buda et al., 2016; Immonen et al., 2014; Jaakkola et al., 2014) and, in these dual roles, are equally impacted by the sociotechnical barriers linked to open data use.

These impediments are encountered along the distinctive phases of providers' as well as consumers' interaction with open data. In a later work, Zuiderwijk et al. (2015) depict corporate activities for commercial open data use: search open data, find, use, and enrich open data, and interpret findings. We also note that governments, as opposed to other open data consumers, undertake steps for publishing open data that resonate with their counterparts' actions in using open data. In the context of data analytics, Hendler (2014) distinguishes between three major steps in the use of heterogeneous online datasets: discovery, integration, and validation. Finally, Abella et al. (2019) suggest that open data reuse, as a concluding step of the proposed open data process, will have a social and economic impact on the surrounding society.

2.4 Research gap

In order to benefit from open data, its consumers (enterprises in particular) must devise efficient approaches to discover and prepare open data for use (Enders et al., 2020). Apart from initial attempts to define open data consumption processes, only a few guidelines assist enterprises in overcoming the main barriers in open data adoption. Open data assessment techniques are one of the ways to tackle the quality-related adoption barriers. Existing efforts predominantly assess open data's metadata quality, rather than the quality of the datasets (Osagie et al., 2017), and largely ignore the use context.

To date, we lack holistic approaches that enable enterprises to efficiently prepare open data for use. A holistic approach would consider the use context and concretize the general steps of finding (identifying), analyzing, and processing (integrating and validating) open data. It would also include methodological guidelines that could help companies overcome the existing barriers (a lack of transparency, heterogeneity, and the unknown quality of open datasets). This endeavor, however, requires integrating fragmented research streams related to open data quality into a more comprehensive approach.

3 Methodology

3.1 Research objectives and setting

Our research aims to develop prescriptive knowledge in the form of a meaningful method to screen, assess, and prepare open data for use in an enterprise setting. It therefore falls under the umbrella of the design science (DS) paradigm, which aims at solving real-world problems and purports to create solutions, often referred to as artifacts, which can take the form of models, constructs, instantiations, or methods (Peppers et al., 2007). Action Design Research (ADR), as a specific DS approach, consists of four main stages, which guide the rigorous process of building artifacts of organizational relevance, and is based on insights gained from practical implementations (Sein et al., 2011). In contrast to existing DS methods that relegate evaluation to a subsequent phase, ADR incorporates evaluation into the design cycles (Sein et al., 2011). It allows to create rigorous and relevant business knowledge that will help to develop “specific solution(s) in specific situation(s)” (Andriessen, 2008) and learn from the instantiations. The outcome of our research is categorized as a method that explains “what to do in different situations” (Goldkuhl et al., 1998) in accordance with a stepwise structure, while also including additional constituents such as notation, procedural guidelines, and concepts (Sandkuhl & Seigerroth, 2019), thereby specifying and documenting the “what” and “how” of the work to be done. It can be considered as a type V theory in terms of Gregor’s (2006) taxonomy of IS research.

Since our artifact purports to solve the problems related to open data identification and preparation for use, the interactions with practitioners are critical for a successful research outcome (Hevner et al., 2004). Our research was conducted in a close industry-research collaborative setting by a team of researchers (two PhD students, two senior researchers, and three master’s students) who worked with a data service provider and data experts from 15 multinational companies. These large multinational companies represent retail, pharmaceutical, automotive, engineering, manufacturing, and chemicals industries.

3.2 Research process

In order to accumulate prescriptive knowledge with the due scientific rigor in an iterative research process, we adhere to the four main stages recommended by Action Design Research (Sein et al., 2011). The first stage of ADR – serving as a starting point to formulate the research effort – is initiated by a problem identified in practice or anticipated by researchers. Among the main activities of this stage, we typically find the initial investigation of the problem, the

determining of its scope, the assignment of roles, and the formulation of the research question(s). In our case, the problem formulation stage debuted in 2017 with several explorative focus groups with practitioners involved in the industry-research collaboration. The primary aim of these focus groups was to identify relevant open data use cases within the companies and to understand their challenges and requirements (see subsection 4.1).

Building on the problem framing and theoretical foundations, the building, intervention, and evaluation (BIE) stage interweaves focus on the design of the artifact. This design is subsequently refined through ongoing organizational use and design cycles, with the process being iterative and taking place within a specific target environment. Table 30 provides an overview of the key elements of the two BIE cycles and the relevant contributions to the development of the method. Our first BIE cycle was part of a multiyear research project (2018-2021) that resulted in a productive platform for data quality services, operated by the data service provider. This platform focuses on business partner curation. Over time, 49 open datasets were onboarded onto the platform (status as of September 2022) to validate and enrich business partner data. In the formalization of learning stage following the first BIE cycle, we aimed to convert the situated learning into general guidelines that support the identification and integration of open datasets. In this phase, the first version of our method was developed based on analyzing the practices that the service provider established to select and prepare datasets and to integrate them with heterogeneous target systems. This version comprises the method's nominal steps and the supporting use of knowledge graphs to explicate business concepts and link them to related datasets. It was evaluated with practitioners during five focus group discussions.

The second BIE cycle was a two-year research project (2019-2021) that aimed to build an open data catalog for business purposes and resulted in a prototype implementation. It encompasses a broader research scope that focuses on an extensive number of use cases, generated in conjunction with the research team and three Swiss-based companies (within telecommunication, public transportation, and fast-moving consumer goods industries), and elaborated on by the data service provider specialists. We applied the method to more than 10 business scenarios (e.g., customs clearance, marketing, and customer analytics) to identify 40 open data use cases, screen and assess relevant open datasets, and map their data models. The discussion of potential use cases for open data led to a systematic approach to use case ideation. Based on our experiences in applying the method to use cases in marketing (e.g., social events and customer targeting), we made several key additions to the different phases, including the development of the assessment phase.

	First BIE cycle	Second BIE cycle:
Context	Development of a productive platform for data quality services, integrating open datasets for validation and enrichment of business partner data	Development of an open data catalog for enterprises (research prototype), that provides open datasets for selected business scenarios
Method development	Alpha version of the method: <ul style="list-style-type: none"> • Development of the method's phases 1 to 3 • Focus on Phase 3 (preparation for use) 	Beta version of the method: <ul style="list-style-type: none"> • Addition of preparatory Phase 0 (use case ideation) • Refinement of phases 1, 2, and 3 in terms of activities and underlying techniques
Main methodological contributions	Phase 3: Knowledge graph to define business concepts, map external datasets, and integrate the datasets into internal systems	Phase 0: Use case ideation approach Phase 2: Three-step assessment comprising metadata, schema, and dataset content level
Evaluation / use cases	Business partner curation, 49 datasets	Ten business scenarios and 46 use cases; assessment of 23 data domains and 220+ datasets

Table 30. BIE cycles and their contribution to method development

In the formalization of learning stage, we reflect on the insights gained from the two BIE cycles, i.e., building of platforms that support companies' use of open data and implement several use cases that are relevant for multinational firms. All steps of the method were fully documented, demonstrated, and additionally discussed in two focus groups with 12 participants from eight companies and 14 participants from 11 companies, respectively. Subsequently, the method was further consolidated, and its separate components (assessment, documentation, and reference ontology model for the selected use cases) were discussed, demonstrated, and evaluated in three individual two-hour sessions with practitioners from the previously mentioned Swiss-based companies. This smaller group of experts are leaders of open data initiatives within their respective companies, and they helped us to better understand the application and usefulness of the suggested method in the enterprise setting. These sessions enabled us to review our design considerations and evaluate our artifact in terms of applicability, consistency, scalability, and understandability criteria (Prat et al., 2015). The sessions were concluded with a questionnaire, through which the method was evaluated by using a five-point Likert scale. Generally, the participants fully agreed (3/3) that the proposed method supports the discovery of the relevant datasets for selected business purposes, agreed (2/3) and fully agreed (3/3) that it supports the assessment and comparison of existing datasets, and agreed (1/3) and fully agreed (2/3) that it supports the mapping of the dataset's attributes to business concepts. They also agreed (1/3) and fully agreed (2/3) that the proposed overall approach to open data integration enables their companies to make better use of open data, and that it could be implemented in their company.

4 Method to screen, assess, and prepare open data for use

4.1 Purpose and design considerations

The method aims to support companies when they identify and prepare suitable open datasets for use in specific business scenarios. It addresses the three issues highlighted in the literature (see subsection 2.1) and confirmed by practitioners during the problem formulation stage: a lack of transparency, heterogeneity, and the unknown quality of open datasets. To provide a systematic and integrated approach, the method design is guided by three important design considerations:

1. *Open data identification should be facilitated and guided by a specific use context that is relevant for the company (screening).* There is a clear need to incorporate the use context in order to identify relevant datasets and understand whether they are “usable for the intended purpose of the user” (Welle Donker & Van Loenen, 2017). Our method suggests goal-oriented, guided search for open data supported by typical use case categories with open data and a structured use case documentation template to capture the relevant internal and external data objects. In contrast to the standardized approaches, it therefore addresses the need for context-aware approaches and assessments (Krasikov, Obrecht, et al., 2021).
2. *The method should help companies gain transparency about relevant datasets and assess their fitness for use (assessing).* To understand whether a candidate dataset is fit for use, the suggested method requires three levels of assessment. Firstly, at metadata level, assessment facilitates the obtainment of primary insights through the description provided at the source level, as suggested by many open data quality assessment techniques. Secondly, at a schema level, assessment is required to determine if the necessary attributes are present within the dataset and whether they will be sufficient to fulfill the use case requirements. This schema-completeness analysis is grounded in the literature on contextual data quality (Pipino et al., 2002; Richard Y. Wang & Diane M. Strong, 1996). Thirdly, at a content level, assessment through traditional data quality metrics is deemed necessary to improve the transparency of the open dataset.
3. *Open data integration needs to consider the existing systems and platforms and map open datasets to internal data models (preparing for use).* Given the heterogeneity of the open datasets and the complexity of their integration, our method relies on knowledge graphs and the concepts of linked data powered by semantic web technologies (Zuiderwijk et al., 2015;

Bizer et al., 2009; Auer et al., 2007; Zaveri et al., 2016). The conceptualization of the domain of interest through ontologies is a known solution when it comes to the integration of large and unknown datasets (Catarci et al., 2017). The use of Ontology-Based Data Access is considered natural when publishing open data, but it requires well-defined semantics of the “right open dataset” (De Giacomo et al., 2018). Our proposed method therefore relies on this common practice for the conceptual mapping of various datasets with identical entities through a graph-based representation of this knowledge, where “the entities, which are the nodes of the graph, are connected by relations, which are the edges of the graph ... and entities can have types, denoted by *is a* relations” (Paulheim, 2016).

4.2 Phases and illustration

The method is structured along four core phases, starting with use case ideation, and thereafter encompassing the screening, assessment, and preparation of open data. Table 32 presents an overview of our method, with each phase having one or more steps, described with goals, main activities, and outcomes. The method comprises techniques and documentation templates (when appropriate) for the introduced steps. In the next subsections, we present each phase with reference to goals, activities and techniques, and practical examples, as well as with reference to the relevant concepts and embedded approaches.

Phase 0 – Use case ideation.

The combination of internal data with open data has proved to be beneficial in different business scenarios (Baud et al., 2002; Schatsky et al., 2019; Strand & Syberfeldt, 2020). Being an initial phase of our method, use case ideation is a mandatory step to understand how open data could complement the enterprise data and help to address specific business problems. Based on our analysis of the business scenarios, we distinguish three generic motivations and use cases with open data: (1) *data management*, i.e., data curation, enrichment, and validation using open reference data, (2) *business processes*, i.e., the improvement of existing processes with the help of externally maintained open data, and (3) *analytics and intelligence*, i.e., the enhancement of analytical insights and predictive models with open data. To define the use case and its context, we propose a template to capture the idea and key notions of the desired use of open data by using four main building blocks: open datasets and providers, data objects (internal and external business concepts/attributes), data management impact, and business impact invoked by the use case. In the early stages of open data initiatives, these notions help to establish the objectives

of open data use and the requirements towards the new data, as they set the scope that enables the screening and assessment activities during the further stages of the proposed method. Building such use cases helps narrowing the scope of the desired open datasets and formulating the selection requirements in the screening phase.

Table 31 illustrates these building blocks for three selected use cases: business partner data curation (an example of a data management use case), customs clearance (an example of a business process use case), and customer analytics (an example of an analytics use case). The template supports the drafting of appropriate potential sources and datasets for the use cases, defining the requirements towards them, and deriving relevant business concepts (or entities) that correspond to the typical attributes of the open datasets.

Use case category and example	Description	Open datasets and providers	Internal data objects	Data management impact	Business impact
<i>Data management use case:</i> Business partner data curation	Leverage open corporate data to increase the quality and knowledge of our business partners (suppliers and consumers)	National corporate registers, global open data company registers (GLEIF, OpenCorporates)	Business partner master data: identification (company name, identifier), address details (country, administrative area, locality, postal code, thoroughfare), and organizational information (data of incorporation, incorporations status, legal form)	Validation of new entries and existing records; Enrichment with new business partner data from open sources; Curation of current business partner data	Prevent billing errors; Automation of data quality activities; Reduced time for data maintenance and entry
<i>Business process use case:</i> Customs clearance	Improve the customs clearance process by using universal standardized codes for product/service classification, tax tariffs, dangerous goods, etc.	World Customs Organization, national customs offices, United Nations, ISO, industry classification (SIC, NACE, EU)	Product data (item name, identifiers, classification, transported quantities, units), commodity codes, and tax tariffs rates	Enrichment of product and supplier data with classification codes; Adherence to international standards; Automation of data maintenance (pre-filled fields)	Reduction of operational cost and customs fees; Improved coordination with customs authorities
<i>Analytics and intelligence use case:</i> Customer analytics	Enhance customer analytics using openly available data provided by public authorities on population, demographics, income, etc.	National statistics office (e.g., Swiss Federal Statistical Office, Eurostat), geographical data (e.g., OpenStreetMap)	Customer data (address), reporting (sales figures and analytics), customer segments	Enrichment of customer data with openly available statistics; New granularity for data analytics	Improved customer outreach; Marketing budget allocation; Improved sales figures

Table 31. Example of use case ideation

	Phase 0	Phase 1	Phase 2			Phase 3	
	0. Use case ideation	1. Identification of relevant open data sources and datasets	2.1 High-level assessment of metadata	2.2 Schema-level assessment	2.3 Dataset content analysis	3.1 Semantic documentation of open datasets	3.2 Integration of open datasets with internal data
Goal	Define and document the use case for open data	Identify relevant sources and underlying datasets	Assess the metadata available at the source	Understand the use case feasibility	Assess the dataset content	Document open datasets	Prepare the datasets for further use by mapping open data with internal data
Main activities	<ul style="list-style-type: none"> Specify the context for which new data is needed in the company Collect potentially relevant sources, decide on relevant business concepts located in open data and their counterparts in internal data Estimate the business impact to concretize the motivation of using open data 	<ul style="list-style-type: none"> Search for and select suitable datasets from open data portals, dedicated search engines, metasearch engines, or expert knowledge of relevant concrete sources Search for authoritative sources or sources that fit the purpose of the use case Define relevant business concepts for the use case as reference ontology 	<ul style="list-style-type: none"> Analyze the metadata provided at source level Check the descriptive statistics of the dataset if available at the source level Verify minimal requirements toward the dataset 	<ul style="list-style-type: none"> Assess schema completeness for the required attributes predefined for the use case Analyze the presence of the required attributes for use case feasibility 	<ul style="list-style-type: none"> Assess content quality based on the applicable data quality dimensions 	<ul style="list-style-type: none"> Provide full metadata documentation, including access, licensing, provenance, and publisher details Document the dataset attributes 	<ul style="list-style-type: none"> Associate the identified attributes with existing business concepts Formulate the mapping and transformation rules for the open data attributes Link open dataset attributes with company entities
Techniques	Use case documentation template	Goal-oriented, guided search for open data	Three-level assessment of open data quality (metadata, schema, and dataset content), combined with traditional data quality dimensions such as completeness, uniqueness, and validity			Documentation and cataloging of open datasets	Knowledge graph to facilitate semantic integration
Outcomes	Documented use cases (based on a template, comprising potential open sources and datasets, as well as business and management impact)	A list with names of datasets, publishers, and data sources A reference ontology	Shortlist of selected datasets	Business concept mapping in the knowledge graph	Decision on which open datasets to be considered for further use in the defined use cases	Detailed dataset documentation	Integrated open datasets

Table 32. Overview of the method to screen, assess, and prepare open data for use

Phase 1 – Screening.

Upon the defined context for open data use, this phase aims to identify suitable data sources and datasets that cover the relevant business concepts for the use case. Open data is available from various providers, such as governments, non-governmental organizations, and companies. While open government initiatives offer access to a large number of open datasets via open data portals (e.g., data.gov, U.S. Census Bureau, or data.europa.eu), some of these open datasets are also discoverable via traditional or dedicated dataset search engines (e.g., Google dataset search or Socrata). In this regard, open data users not only have to identify relevant datasets but must also verify the authoritativeness (publisher details) of the source by means of the provided metadata, if available. The absence of such information raises concerns about the source and content of the underlying data.

For the use case of business partner data curation, Table 33 presents examples of identified datasets for corporate registers from leading EU countries in the open data initiatives (van Hesteren et al., 2022) and leading world economies with recognized open data initiatives (Global Open Data Index, 2015; OpenDataBarometer & World Wide Web Foundation, 2020), along with the acknowledged data sources and publisher information. Only publicly available datasets provided in downloadable and machine-readable formats were considered. It is important to note that for corporate registers, multiple sources lead to the desired dataset, e.g., crawled search engines like Google dataset search or open data initiatives like Global Legal Entity Identifier Foundation's (GLEIF). In this regard, GLEIF aggregates, registers, and currently lists more than a thousand corporate registers across the world (GLEIF, 2019), which are provided by official authorities. It thereby provides a link to sources that are often deemed authoritative since they are published and maintained by competent governmental agencies (e.g., the state/government departments or ministries).

In this phase, the previously identified business concepts (see Phase 0) can be extended with concepts derived from open datasets. They represent the reference ontology that can be used for concept mapping and specification of relationships between internal business objects and the open datasets, in line with the knowledge graph principles.

Dataset	Publisher	Sources
Argentinian National Registry of Companies	Ministry of Justice and Human Rights (Argentina)	Argentina.gob.ar, GLEIF, Google dataset search
Colorado Business Entity Register	Colorado Department of State	Data.colorado.gov, data.gov, GLEIF
French Register of Companies	National Institute of Statistics and Economics Studies (France)	Sirene.fr, GLEIF
Latvian Register of Enterprises	The Register of Enterprises of the Republic of Latvia	Dati.ur.gov.lv, GLEIF
Norwegian Register of Business Enterprises	The Central Coordinating Register for Legal Entities	Data.brreg.no, GLEIF
New York Business Entity Register	New York Department of State	Data.ny.gov, data.gov, GLEIF
UK Companies House	Companies House (UK)	Gov.uk, Google dataset search, GLEIF

Table 33. Example of identified open data sources and datasets

Phase 2 – Assessment.

During this phase, candidate datasets are analyzed to determine their suitability for the defined use case. The underlying process for Phase 2 is threefold and is conducted on the metadata, schema, and content levels of the datasets. By providing a context-specific assessment of a dataset's schema and content, it thereby extends beyond the existing open data assessment approaches presented in subsection 2.2. Each of the subphases is accompanied by specific criteria that may lead to the selection or rejection of a dataset. The sequential assessment (metadata – schema – dataset content) helps to preselect relevant datasets on the metadata level, minimizing the risk of wasted efforts on datasets with unclear content, which is particularly relevant in the enterprise setting. We argue that to understand the open data's "usability", an analysis of the use case-specific attributes must be incorporated along with the traditional assessment approaches.

To formalize the content-aware assessment phase of our method, we consider the relevant dimensions and metrics (see Table 34), suggested by the comprehensive approaches of Neumaier et al. (2016), Vetrò et al. (2016), and Zhang et al. (2019). As discussed in subsection 2.2, although open data assessment approaches build on traditional data quality dimensions, they should also consider the use context that is relevant and feasible from the practitioners' perspective. Thus, our content-aware selection embodies both perspectives and allows the selection of dimensions that can realistically be assessed in the context of unknown datasets, as indicated by practitioners. Completeness (in its different forms) appears to be one of the most applicable dimensions in open data assessment (Vetrò et al., 2016; R. Zhang et al., 2019), being a primary indicator of whether a dataset can actually be used for the intended purpose. This is largely due to the fact that the absence of the necessary information cannot be easily compensated by traditional data quality improvement approaches (Batini et al., 2009). From the perspective of practitioners, it is often pointless to analyze a dataset which is critically

incomplete or even empty, especially if mandatory attributes, defined as “business concepts” in Phase 1, are not present. While completeness is the dominant dimension at metadata and schema levels, additional dimensions should be included at the dataset content level. Dimensions that can be realistically assessed at the dataset level, besides completeness, are uniqueness (rows) and validity (format compliance).

Subphase	Dimension	Scope	Metric	Description
2.1 Metadata assessment	Metadata completeness	Metadata	Presence or absence of the required metadata entries (at the source level)	Indicates the presence of metadata attributes necessary for the proper identification of the dataset: general information (format, access login, lookup service), licensing presence, publishing details (publisher, publishing date, update cycle), and content-related information (resource language, geographic coverage, number of records, and number of diverse attributes).
2.2 Schema assessment	Schema completeness	Schema	Presence or absence of the required attributes	Represents the degree to which attributes are present in the schema of the dataset. This primarily refers to the relevant fields or attributes of the specific use case.
2.3 Dataset content assessment	Overall cell completeness	Dataset	Percentage of missing cells in the whole dataset	Indicates the percentage of missing cells in a dataset, meaning that the cells that are empty and do not have an assigned value.
	Row uniqueness	Dataset / record	Percentage of duplicate rows	The data record is uniquely identifiable.
	Completeness of mandatory attributes	Dataset / column	Percentage of missing cells within a column	The attributes which are mandatory for a complete representation of a real-world entity must contain values and cannot be null. This can also include the mandatory attributes of the predefined use case, based on the requirements.
	Metadata compliance / understandability	Dataset / column	Percentage of compliant cells within a column	The data should comply with its metadata. It indicates the percentage of cells within a column in a dataset that complies with metadata specifications.
	Format compliance	Dataset / column	Percentage of compliant cells within a column	Indicates the percentage of cells within a column that comply with the format specified for the column in a dataset. It only considers the columns that represent some kind of information associated with standards (e.g., geographic information).

Table 34. Relevant open data quality dimensions on metadata, schema and dataset content level, based on Neumaier et al. (2016), Vetrò et al. (2016), and Zhang et al. (2019).

Subphase 2.1. This subphase begins with a high-level analysis of metadata, typically available at the source level, which is the focus of most of the open data assessment methods. Neumaier et al.’s (2016) metadata quality assessment framework suggests the verification of the existence of metadata attributes of Data Catalog Vocabulary (DCAT) (Albertoni et al., 2020) as a W3C metadata recommendation for publishing data on the Web. Although the approach itself is suitable for the necessary level of assessment and the commonly used completeness metric

(Pipino et al., 2002), current DCAT metadata attributes do not cover all of the attributes identified in our research process. As the minimal information related to identifying a dataset, we consider metadata attributes describing the access conditions (format, access login, lookup service), licensing presence, publishing details (publisher, publishing date, update cycle), and general content-related information (resource language, geographic coverage, number of records, and number of diverse attributes). With this information at hand, simple rejection criteria can be verified (e.g., no access to the data, no machine-readable formats, non-open license). Violating these criteria will lead to the dataset being removed from further investigation. If available, descriptive statistics of the datasets' contents can also be considered at the source level, for example the number of downloads, ratings, and number of rows and attributes in a dataset, as well as the file size.

Subphase 2.2. Upon completing the metadata assessment, an initial investigation can be done into the datasets, starting with their data model. This schema-level assessment ensures that the required attributes for the use cases are present in the dataset and that the dataset is “usable” (Krasikov, Obrecht, et al., 2021). For this purpose, the completeness of each dataset's schema is further analyzed, allowing a verification of the presence of the mandatory attributes, defined as “business concepts” in Phase 1 through the underlying reference ontology design. This assessment can be conducted using the completeness dimension, “which is the degree to which entities and attributes are not missing from the schema” (Pipino et al., 2002). This step is crucial to understand whether each dataset's content is sufficient to realize the use case, and to comprehend if it is possible to establish the mapping of the concepts present in internal and external datasets. For instance, datasets from corporate registers contain information about enterprises' identification codes and address details (Table 35), but the availability of additional attributes (e.g., company's legal form, activity status, or postal codes) depend on the specific dataset and source.

Subphase 2.3. To finalize the assessment and solidify the selection of open datasets for the use case, it is necessary to conduct a thorough assessment of their content. This assessment focuses on the content of datasets in terms of typical data quality dimensions, such as completeness, uniqueness, validity, and the related metrics. Such approaches are covered in the literature (Vetrò et al., 2016; R. Zhang et al., 2019), but must be adapted for the domains of open datasets in different use cases. To ensure the usability of open datasets for a specific use case, we specifically suggest considering completeness of the mandatory attributes (which are first derived in Phase 0 and then defined as reference ontology in Phase 1). We also consider

uniqueness and validity in this step, as seen in Table 34. After the assessment, a final decision can be made on the suitability of the open dataset for the intended use case.

To illustrate this phase in a real scenario, we exemplify the three-level assessment (i.e., metadata, schema, and content), for seven corporate registers (see Table 33) and the business partner data curation use case, as formulated in Phase 0. To perform these activities, we used the Pandas Profiling (YData Labs, 2023) library, which provides an easy-to-use interface to summarize the various aspects of the datasets. This library's main function `profiling.ProfileReport()` takes a Pandas DataFrame as its input and returns a `ProfileReport` object, which can be rendered as an HTML report. While the report provides suitable grounds for retrieval of the descriptive statistics of the dataset, it lacks depth in terms of use context. Therefore, in this case, it was only used as a supporting tool to perform the calculations. For instance, a section of the report provides a description of variables (i.e., attributes of the dataset), including the variable types, number of unique values, missing values, and distribution of values. In this example, metadata-level and schema-level assessments were performed manually, even though this process can be automated in productive implementation. To illustrate the dataset content assessment, we extracted the values from the profiling report and demonstrated the completeness and uniqueness of the corporate registers' datasets (see Table 35). For demonstration purposes, the names of the actual attributes within the datasets were renamed to match the reference ontology design, illustrated by the next phase of the method. We also provided observations for each dataset (see first column of Table 35). An important shortcoming of automatic profiling tools is the verification of the presence of the mandatory attributes, the definition of which is based on the use case requirements (see Phase 0). The underlying reference ontology design helps to identify these mandatory data objects within open datasets, based on the internal data objects (in the event they are known) or defined as "business concepts" in Phase 1.

This example reveals the particularities of the assessment phase, especially in terms of conclusions drawn about the datasets, based on the three suggested pillars. For instance, the overall completeness of the datasets (total missing cells %) is not unfavorable for the use case. However, from a traditional assessment perspective, this incompleteness is often interpreted as constituting poor data quality. The prominence of this deficiency becomes even more noticeable when dealing with large datasets as a large-scale automated assessment would flag the high number of missing values. In our case, more than half of all cells were missing in several company registers (e.g., in the UK and France); however, the individual completeness of mandatory attributes can render the dataset usable for the formulated use case. To the contrary,

we similarly note that the individual completeness of the mandatory attributes should be regarded with caution. If the attribute in question contains an alarming number of missing cells, the whole dataset could be deemed unusable for the use case. When dealing with uniqueness, the identifier attributes for the assessed corporate registers help us to cope with possibility of duplicate rows and, in the case of this analysis, the assumed authoritativeness and rigor of the governments data help us to keep track of the registered companies in a standardized manner within given legislation.

Dataset	Metadata	Schema	Content
Argentinian National Registry of Companies <i>Observations: The dataset is published with clear access details, all mandatory attributes are present, and there is a low percentage of missing values in the specific attributes of the use case.</i>	Identification: RA000010 Country: Argentina Format: CSV Access login: no Free lookup service: available License: Creative Commons Attribution 4.0 Publishing date: 19.09.2016 Update cycle: 30d Geographic coverage: National # of records: 1'057'485 # of attributes: 22	10/10 mandatory attributes for the use case of "Business partner data curation"	Total missing cells (all dataset): 18.7% Total duplicate rows (all dataset): 0.0% Attribute (company name): 0.0% missing Attribute (identifier): 0.0% missing, 100% distinct Attribute (country): 0% missing Attribute (administrative area): 2.9% missing Attribute (locality): 2.9% missing Attribute (post code): 2.9% missing Attribute (thoroughfare): 2.9% missing Attribute (legal form): 2.4% missing Attribute (status): 2.9% missing Attribute (date of incorporation): 1.1% missing
Colorado Business Entity Register <i>Observation: The dataset is well-published with clear metadata and necessary attributes to determine use case feasibility, but overall incompleteness on the attribute level renders it unusable.</i>	Identification: RA000599 Country: United States Format: CSV, RDF, RSS, TSV, XML, REST Access login: no Free lookup service: available License: Public Domain Publishing date: 19.03.2014 Update cycle: 1d Geographic coverage: State # of records: 1'048'575 # of attributes: 35	10/10 mandatory attributes for the use case of "Business partner data curation"	Total missing cells (all dataset): 84.7% Total duplicate rows (all dataset): 0.0% Attribute (company name): 79.4% missing Attribute (identifier): 0.0% missing, 99.9% distinct Attribute (country): 79.9% missing Attribute (administrative area): 79.9% missing Attribute (locality): 79.9% missing Attribute (post code): 79.9% missing Attribute (thoroughfare): 79.9% missing Attribute (legal form): 79.4% missing Attribute (status): 79.4% missing Attribute (date of incorporation): 79.4% missing
French Register of Companies <i>Observation: Given the size of the dataset, it is well-published, but its overall completeness is less than 50%, even though the individual completeness of the mandatory attributes enhances its usability.</i>	Identification: RA000189 Country: France Format: CSV, API Access login: no Free lookup service: available License: Open License V2.0 Publishing date: 24.08.2018 Update cycle: 1d Geographic coverage: National # of records: 32'648'533 # of attributes: 48	9/10 mandatory attributes for the use case of "Business partner data curation"	Total missing cells (all dataset): 59.7% Total duplicate rows (all dataset): 0.0% Attribute (company name): 92.8% missing Attribute (identifier): 0.0% missing, 100% distinct Attribute (country): 0% missing Attribute (administrative area): 0.8% missing Attribute (locality): 81.3% missing Attribute (post code): 0% missing Attribute (legal form): 0% missing Attribute (status): 0% missing Attribute (date of incorporation): 1.6% missing
Latvian Register of Enterprises <i>Observation: Although certain</i>	Identification: RA000423 Country: Latvia Format: CSV, XSLX Access login: no	10/10 mandatory attributes for the use case of "Business	Total missing cells (all dataset): 13.9% Total duplicate rows (all dataset): 0.0% Attribute (company name): 0.1% missing Attribute (identifier): 0.0% missing, 100% distinct

Dataset	Metadata	Schema	Content
<i>details are missing in the metadata, the dataset is maintained with a comparably high level of quality.</i>	Free lookup service: available License: n/a Publishing date: 10.03.2014 Update cycle: n/a Geographic coverage: National # of records: 440'422 # of attributes: 21	partner data curation"	Attribute (country): 0% missing Attribute (administrative area): 0% missing Attribute (locality): 0% missing Attribute (post code): 4.6% missing Attribute (thoroughfare): 0.1% missing Attribute (legal form): 0% missing Attribute (status): 0% missing Attribute (date of incorporation): 0.1% missing
Norwegian Register of Business Enterprises <i>Observation: The metadata lacks several important entries and even though the mandatory attributes are present, the address information is absent in approximately 80% of the values.</i>	Identification: RA000472 Country: Norway Format: CSV, JSON, XML, REST, API Access login: no Free lookup service: available License: Norwegian Open License Publishing date: n/a Update cycle: n/a Geographic coverage: National # of records: 1'048'575 # of attributes: 43	10/10 mandatory attributes for the use case of "Business partner data curation"	<u>Total missing cells (all dataset): 37.2%</u> <u>Total duplicate rows (all dataset): 0.0%</u> Attribute (company name): 0% missing Attribute (identifier): 0.0% missing, 100% distinct Attribute (country): 80.7% missing Attribute (administrative area): 81.6% missing Attribute (locality): 80.7% missing Attribute (post code): 81.6% missing Attribute (thoroughfare): 80.8% missing Attribute (legal form): 0% missing Attribute (status): 0% missing Attribute (date of incorporation): 0.8% missing
New York Business Entity Register <i>Observation: Although the dataset is accessible, its overall completeness is less than 50% and it lacks two mandatory attributes. The present attributes are, however, complete.</i>	Identification: RA000628 Country: United States Format: CSV, RDF, RSS, TSV, XML Access login: no Free lookup service: available License: Open Government Publishing date: 14.02.2013 Update cycle: 30d Geographic coverage: State # of records: 3'308'768 # of attributes: 30	8/10 mandatory attributes for the use case of "Business partner data curation"	<u>Total missing cells (all dataset): 54.0%</u> <u>Total duplicate rows (all dataset): 0.0%</u> Attribute (company name): 0.0% missing Attribute (identifier): 0.0% missing, 100% distinct Attribute (country): 0.5% missing Attribute (administrative area): 2.2% missing Attribute (post code): 2.5% missing Attribute (thoroughfare): 2.2% missing Attribute (legal form): 0% missing Attribute (date of incorporation): 0% missing
UK Companies House <i>Observation: A well-published dataset that includes all mandatory attributes. While the level of overall completeness is insufficient, most of the mandatory attributes are complete.</i>	Identification: RA000585 Country: United Kingdom Format: CSV, REST Access login: no Free lookup service: available License: Open Government v3.0 Publishing date: 11.12.2016 Update cycle: 7d Geographic coverage: National # of records: 5'063'321 # of attributes: 55	10/10 mandatory attributes for the use case of "Business partner data curation"	<u>Total missing cells (all dataset): 50.9%</u> <u>Total duplicate rows (all dataset): 0.0%</u> Attribute (company name): 0% missing Attribute (identifier): 0.0% missing, 100% distinct Attribute (country): 0.0% missing Attribute (administrative area): 65.3% missing Attribute (locality): 1.8% missing Attribute (post code): 1.3% missing Attribute (thoroughfare): 0.9% missing Attribute (legal form): 0% missing Attribute (status): 0% missing Attribute (date of incorporation): 0.0% missing

Table 35. Examples of the datasets' assessment results on the metadata, schema, and content levels

Phase 3 – Preparation for use.

This phase entails the integration of the identified and assessed open datasets in a company's internal system. The identified business concepts and reference ontology are key for the concept

mapping and specification of relations between the entities, in line with the knowledge graph principles. Semantic technologies provide a more robust and flexible way of integrating data from multiple sources, because they use a common vocabulary and data model, which facilitate the linkage and integration of data obtained from different sources (Van Nuffelen et al., 2014). In addition, semantics can also improve the quality of the integrated data by allowing data validation and reconciliation that use ontologies and formal logic. This can ensure the accuracy and consistency of the integrated data, which is particularly important when dealing with open data that may have been collected by different organizations or from different sources.

Subphase 3.1. Our method recommends a thorough documentation of the selected open datasets and the provision of complete metadata information. Certain open data sources (e.g., open data portals) already adhere to well-known metadata vocabularies and standards (e.g., DCAT, DCT, DQV, SDO), which simplify the documentation process by having standardized RDF vocabularies for metadata description. A common metadata model for the documentation of open datasets assists their harmonization, increases transparency, and documents additional aspects such as quality and dataset attributes. In addition, the documentation of attributes should contain the associated business concepts (as seen in Phases 0 and 1), as this allows to initiate the construction of the knowledge graph.

Subphase 3.2. This final subphase focuses on integrating open datasets by means of a knowledge graph. The previous subphase emphasized the links between open dataset attributes and the common entities (business concepts), thus denoting the formalization of an ontological model for a given use case. For instance, a company's internal data objects need to be associated with similar entities as those found in open datasets. This entity-linking process is a common way of integrating heterogeneous datasets (Zuiderwijk et al., 2015; Bizer et al., 2009; Auer et al., 2007; Zaveri et al., 2016). As a result, a company will be able to locate open datasets containing attributes that correspond to business concepts, which in turn relate to their internal data.

To illustrate these subphases in a real scenario, we once again refer to the data management category of the use cases, namely business partner curation. As suggested in subphase 3.1, a thorough documentation of open datasets is necessary to prepare the concept linkage and, as part of the first BIE cycle (see Table 30), a productive documentation is maintained using a MediaWiki with an extension of Semantic MediaWiki. Figure 12 provides an example of an open dataset's metadata documentation on a web-based semantic engine, for example, the commercial register of France (see https://meta.cdq.com/Data_source/FR.RC), including the metadata of the dataset, as well as its attributes, concept mappings, values, and value mappings.

This documentation informs the open data consumer, thereby improving the transparency of the dataset’s provenance, as well as of its content. On a more abstract level, since several datasets can be linked to the same concepts, e.g., the New York Business Entity Register contains mandatory attributes that matches those of the French Register of Companies (see Figure 13).

Page Discussion

Data source FR.RC

Data source|FR.RC

Name	ALEI prefix : FR.RC , Full name (en) : France Register of Companies, Short name : France Company Register, Local name : Répertoire administratif SIRENE
Technical key	FR_RC
Short description	Commercial register of France.
Description	Allows to verify the validity of French business identifiers (SIREN number (France), SIRET number (France)) and gather additional information about the company status, registered address, and names.
Country scope	FR (Frankreich, France, France, 法国)
Provider	Institut national de la statistique et des études économiques, Search type : FUZZY, website
Release status	LIVE , Integration type : COPY
Content and data model	Category : BUSINESS_PARTNER, Number of records : 34,128,160 records (last update : 2023-01-11), 72 concepts, 25 values, 44 mappings, 0 transformations
Linkage strategy	SIRET number (France)
Managed concepts	See mapped concepts [Expand]
Managed identifier(s)	See mapped identifiers [Expand]
Augmented records	See how data source is used [Expand]
Update monitoring	See details of update monitoring [Expand]
Classification	Data ownership : GOVERNMENTAL , Data provision : OPEN , Terms type : EU_PUBLIC , legal approval: 2020-01-01 Licence URL : https://www.sirene.fr/sirene/public/static/conditions-generales-utilisation

Figure 12. Example of the documentation of an open dataset (from https://meta.cdq.com/Data_source/FR.RC)

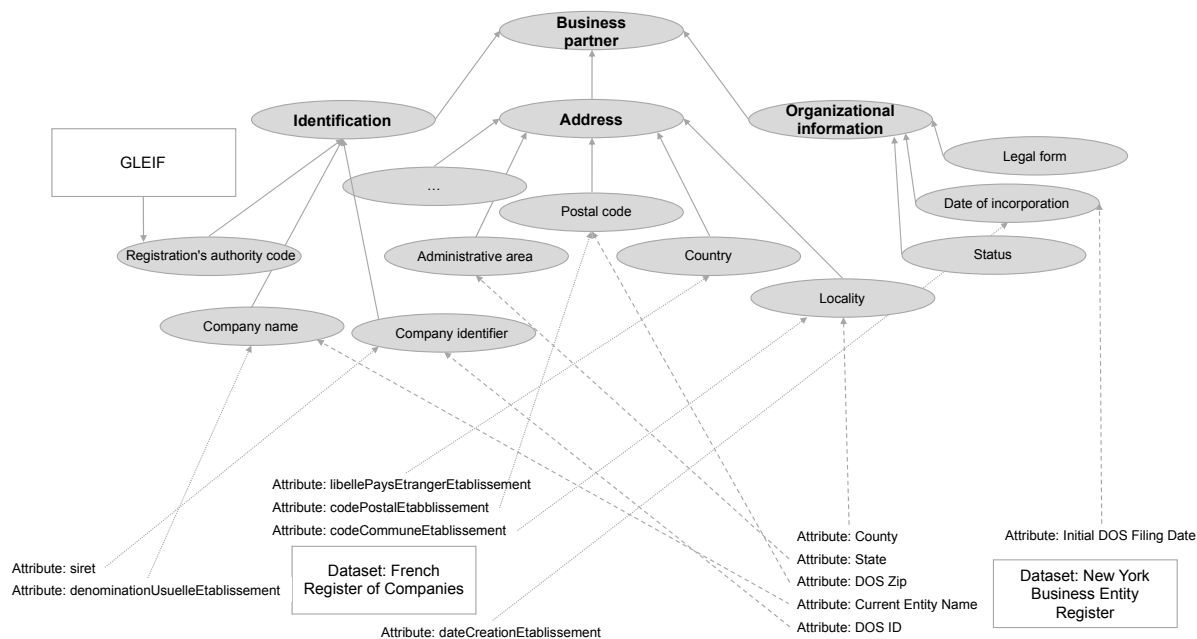


Figure 13. Example of reference ontology and entity-linking process for selected datasets

4.3 Workflow

While our method outlines a systematic approach, covering the phases from use case ideation to open data preparation for use, the method’s application in practice can be non-linear. To illustrate, Figure 14 presents a workflow and thereby highlights the variations and sequencing

of applications of our method in enterprises beyond the main flow illustrated in the previous sections, thus allowing for process flexibility and adaptability.

One of the possible variations concerns the entry points for the suggested method, depending on the situational context of the company. For instance, if the open data use case already exists in the enterprise context, the company can begin with Phase 1, thus starting directly with the identification of relevant open data sources and datasets. Furthermore, if the datasets are already identified, a possible entry point is Phase 2, implying the assessment of the pre-selected datasets. It is necessary to mention that if the enterprise already productively uses open data in defined use cases, revisiting different steps of our method can help rethinking the adopted approach with the intention of improving current practices.

The presented workflow defies the linearity of our method, particularly concerning the assessment phase. As described in subsection 4.2, the quality of the metadata, schema, and content of the open datasets may differ and may potentially not meet the assessment criteria, e.g., when no or not enough datasets are shortlisted. This implies returning to Phase 1 and initiating a new search for suitable datasets, thus repeating the Phases 1 and 2. This variant also occurs when new datasets appear or if there are previously omitted datasets. Additionally, it is possible that the use case requirements as such must be redefined in order to identify suitable datasets. This furthermore implies the adoption of an iterative approach to the assessment of datasets, i.e., revisiting the three levels to ensure a sufficient level of underlying quality.

Finally, even if open datasets have been successfully prepared and integrated for use, there are what-to-do-next options. By going beyond the method's, it is possible that whenever the dataset is updated, the organization could return to Phase 2 to check for any changes (e.g., if the metadata is acceptable, that the schema and content are of a sufficient quality before integrating the updated dataset and existing assets). Upon the successful integration of the open datasets, it is possible that additional use cases may be developed on this basis, thus returning to use-case ideation. Ultimately, since our method does not impose processual steps, it might, in specific contexts, not be necessary to repeat each phase.

A Method to Screen, Assess, and Prepare Open Data for Use

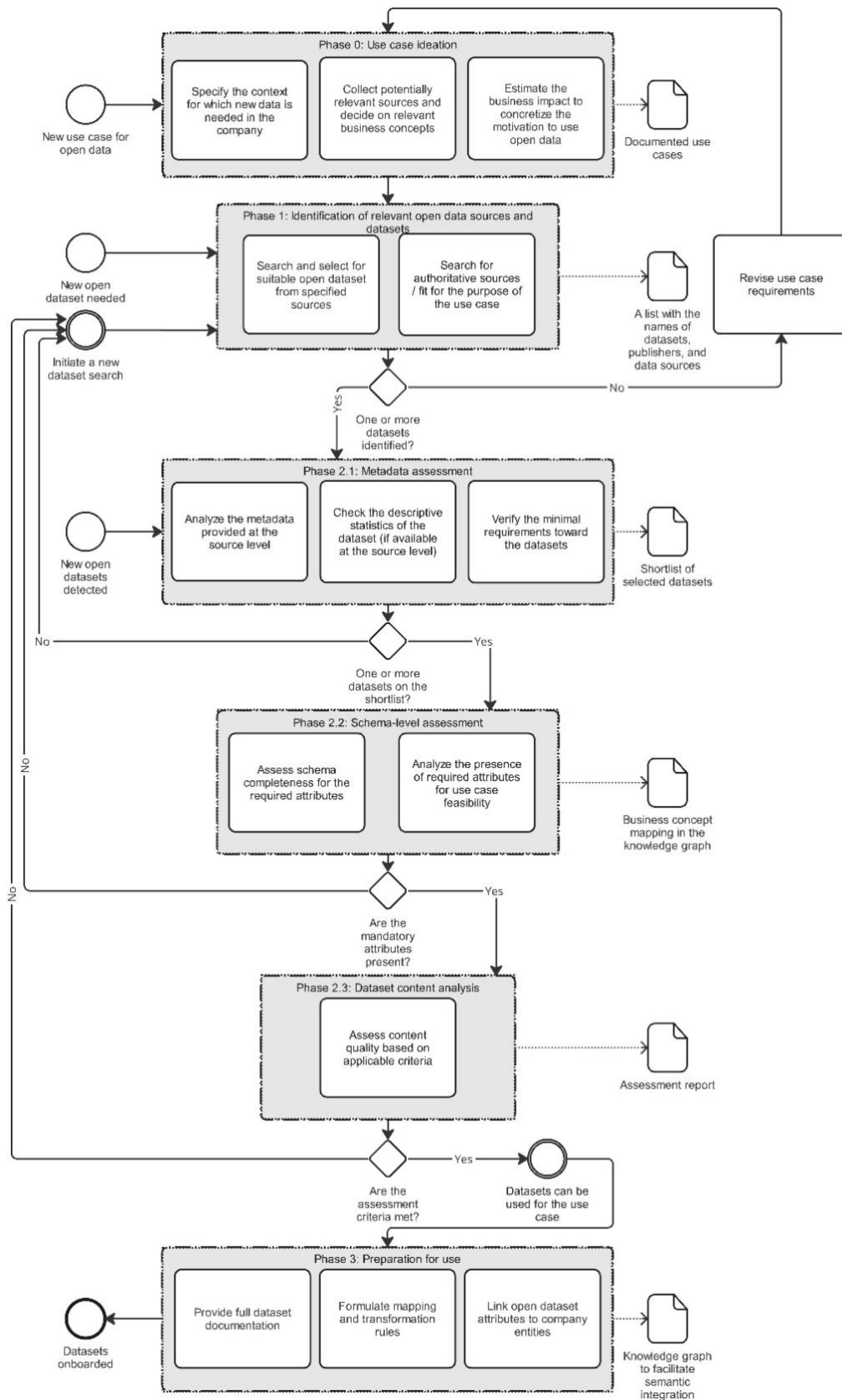


Figure 14. Possible variations in the workflow of the method to screen, assess, and prepare open data for use

5 Comparison with other frameworks and approaches

In order to position our method to screen, assess, and prepare open data for use within the existing body of literature on open data, we return to the existing open data frameworks and approaches discussed in the prior research section (see subsections 2.2 and 2.3). Therefore, Table 36 summarizes how our method compares with existing open data approaches in terms of the design considerations formulated in subsection 4.1. In line with our review in subsection 2.2, in this comparison we consider papers that formulate open data approaches or models and that go beyond the mere presentation of quantitative results.

Approach	Purpose	Screen	Assess		Prepare for use (integration)
			Use context awareness	Scope	
“Luzzu” framework (Debattista et al., 2016)	Linked data quality assessment	no	no	Metadata and dataset	yes
Metadata quality assessment framework (Neumaier et al., 2016)	Automated metadata quality assessment for various open data portals	no	no	Metadata	no
Measurement framework (Vetrò et al., 2016)	Quantitative assessment of open government data quality	no	no	Metadata and dataset	no
Benchmarking framework (Máchová & Lněnička, 2017)	Evaluation of open data portals’ quality	no	no	Metadata	no
Holistic open data assessment framework (Welle Donker & Van Loenen, 2017)	Assessment of the quality of open data supply, open data governance, and user perspective of the open data infrastructure	no	yes	Metadata	no
Quality-based selection framework (Stróżyńska et al., 2018)	Selection of open data sources to be fused with internal data	yes	yes	Metadata	yes
“LANG” approach (R. Zhang et al., 2019)	Discovery of data quality problems in repurposed datasets	no	no	Metadata and dataset	no
Method to screen, assess, and prepare open data for use	Prepare open data of uncertain quality for use in a value-adding and demand-oriented manner	yes	yes	Metadata, schema, dataset content	yes

Table 36. Comparison with other approaches

While the existing approaches and frameworks do have advantages when it comes to an in-depth immersion into the quality aspects, they do not holistically inform and demonstrate how open data can be screened, assessed, and prepared for use in the enterprise context. It is worth

noting that Stróżyńska et al.'s (2018) quality-based selection framework is the only approach which actually covers the screening, assessment (also considering the use context), and preparation for use phases. However, it considers only the open data sources' metadata and primarily covers the selection aspect. Other than that, Welle Donker et al.'s (2017) holistic open data assessment framework is another approach that covers the specific use context for open data. However, it does not provide any guidance on screening and integration aspects.

In terms of assessment techniques, it is important to mention Neumaier et al.'s (2016) metadata quality assessment framework, Vetrò et al.'s (2016) measurement framework, and Zhang et al.'s (2019) "LANG" approach, that were used in Phase 2 (see subsection 4.2), as they all provide a basis for open data quality assessment dimensions on both the metadata level and the dataset level. While the three frameworks do not provide consistent evidence on how to screen and integrate open data (particularly in the context of semantics), it is important to state that this was not their original intention. The common idea of these frameworks and their respective approaches is to standardize and clarify the data quality assessment techniques for external datasets from open sources. None of them claim to provide a holistic approach to open data sourcing.

6 Conclusion and limitations

While the potential of open data is well-known to the research community and to practitioners, the widespread use of open data still lags. In our multiyear research project, we attempted to resolve the main challenges faced by enterprises when engaging in open data use cases related to data management, business processes, or analytics. Our research activities result in a method that comprises four phases and that supports companies through all steps ranging from deciding on the suitable use cases for open data to preparing open datasets for actual use. To the best of our knowledge, this is one of the first systematic attempts to provide methodological guidance to prepare open data of uncertain quality for use in a value-adding and demand-oriented manner.

Compared to prior literature, our method consolidates different streams of open data research by adopting a systematic approach. First, it contextualizes open data use by providing guidance to use case ideation and by exemplifying the generic business scenarios which allow gaining value from open data. It thereby ensures that open data is "usable for the intended purpose of the user" (Welle Donker & Van Loenen, 2017). Second, our method proposes a context-aware open data assessment approach that comprises metadata-, schema-, and content-level

techniques. It thereby reflects open data quality assessment approaches and links them to traditional data quality literature. Third, our method is enabled by the use of semantic concepts for data integration – a knowledge graph and reference ontologies – that allow the mapping of open datasets by linking them to internal data objects. This approach enables enterprises to locate open datasets containing attributes that correspond with business concepts, which in turn relate to their internal data. Our method therefore provides a scalable approach to the integration of heterogeneous datasets (Zuiderwijk et al., 2015; Bizer et al., 2009; Auer et al., 2007; Zaveri et al., 2016).

Our method contributes to practice and research. For practitioners, it goes beyond the existing nominal process steps and outlines a systematic approach with concrete goals, activities, techniques, and outcomes. Therefore, it should be considered as an important pillar of an open data strategy (Enders et al., 2020). For academics, our research conceptualizes open data preparation as a purposeful and value-creating process. Furthermore, our method to screen, assess, and prepare open data for use can not only facilitate the allocation of related research activities along the process chain, but also assist the building of a foundation for future research on specific use cases and open datasets. We strongly believe that our method addresses the research gap related to a lack of elaborate processes for open data use and mechanisms for enterprise-wide open data strategy implementation (Enders et al., 2020). The suggested method also demonstrates how semantic technologies, resulting from technical open data research streams, can be systematically applied and how they can complement organizational processes for open data assessment and use. While the screening and assessment phases of the method are widely applicable, the preparation for use with semantic technologies requires long-term investments. The last phase will require organizations to train their staff in the use of new tools, languages, and methodologies for data integration, management, and analysis.

Nevertheless, this work is subject to limitations. Our specific research context, namely the ADR research project, may limit the generalizability of our findings and the versatility of our proposed method. More specifically, even though our method synthesizes practitioner knowledge garnered from various open data use cases and firms, additional large-scale demonstrations and further evaluations would be beneficial. Since our method comprises context-specific elements, it would benefit from pre-existing reference ontologies for specific business contexts. This offers noteworthy potential for future design science research in the information systems field, namely semantic modeling, and knowledge graphs for open data use.

7 References

- Abella, A., Ortiz-de-Urbina-Criado, M., & De-Pablos-Heredero, C. (2019). The Process of Open Data Publication and Reuse. *Journal of the Association for Information Science and Technology*, 70(3), 296–300.
- Abida, R., Belghith, E. H., & Cleve, A. (2020). An End-to-End Framework for Integrating and Publishing Linked Open Government Data. *Proceeding of the 29th International Conference on Enabling Technologies*, 257–262.
- Albertoni, R., Browning, D., Cox, S., Beltran, A. G., Perego, A., & Winstanley, P. (2020). *Data Catalog Vocabulary (DCAT)*. <https://www.w3.org/TR/vocab-dcat/>
- Andriessen, D. (2008). Combining Design-Based Research and Action Research to Test Management Solutions. In *Towards Quality Improvement of Action Research: Developing Ethics and Standard* (pp. 125–134). Brill.
- Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). DBpedia: A Nucleus for a Web of Open Data. *The Semantic Web*, 722–735.
- Bachtiar, A., Suhardi, & Muhamad, W. (2020). Literature Review of Open Government Data. *Proceedings of the 2020 International Conference on Information Technology Systems and Innovation*, 329–334.
- Barry, E., & Bannister, F. (2014). Barriers to Open Data Release: A View from the Top. *Information Polity*, 19(1/2), 129–152.
- Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 41(3), 1–52.
- Baud, N., Frachot, A., & Roncalli, T. (2002). Internal Data, External Data and Consortium Data—How to Mix Them for Measuring Operational Risk. *SSRN Electronic Journal*, 1–18.
- Beno, M., Figl, K., Umbrich, J., & Polleres, A. (2017). Open Data Hopes and Fears: Determining the Barriers of Open Data. *Proceedings of the 2017 Conference for E-Democracy and Open Government*, 69–81.
- Bicevskis, J., Bicevska, Z., Nikiforova, A., & Oditis, I. (2018). Data Quality Evaluation: A Comparative Analysis of Company Registers' Open Data in Four European Countries. *Communication Papers of the 2018 Federated Conference on Computer Science and Information Systems*, 17, 197–204.
- Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data—The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3), 1–22.
- Bogdanović-Dinić, S., Veljković, N., & Stoimenov, L. (2014). How Open Are Public Government Data? An Assessment of Seven Open Data Portals. In *Measuring E-government Efficiency* (Vol. 5, pp. 25–44). Springer.
- Braunschweig, K., Eberius, J., Thiele, M., & Lehner, W. (2012). The State of Open Data Limits of Current Open Data Platforms. *Proceedings of 21st World Wide Web Conference*.
- Buda, A., Ubacht, J., & Janssen, M. (2016). Decision Support Framework for Opening Business Data. *Proceedings of the 16th European Conference on E-Government*, 29–37.
- Catarci, T., Scannapieco, M., Console, M., & Demetrescu, C. (2017). My (Fair) Big Data. *Proceedings of the 2017 IEEE International Conference on Big Data*, 2974–2979.
- Conradie, P., & Choenni, S. (2014). On the Barriers for Local Government Releasing Open Data. *Government Information Quarterly*, 31, S10–S17.
- Corsar, D., & Edwards, P. (2017). Challenges of Open Data Quality: More Than Just License, Format, and Customer Support. *Journal of Data and Information Quality*, 9(1), 1–4.
- Crusoe, J., & Melin, U. (2018). Investigating Open Government Data Barriers: A Literature Review and Conceptualization. In *Electronic Government* (Vol. 11020, pp. 169–183). Springer.
- Data.gov. (2022). Data.Gov. <https://www.data.gov/>
- De Giacomo, G., Lembo, D., Lenzerini, M., Poggi, A., & Rosati, R. (2018). Using Ontologies for Semantic Data Integration. In *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years* (Vol. 31, pp. 187–202). Springer.
- Debattista, J., Auer, S., & Lange, C. (2016). Luzzu—A Methodology and Framework for Linked Data Quality Assessment. *Journal of Data and Information Quality*, 8(1), 1–32.
- Enders, T., Benz, C., & Satzger, G. (2021). Untangling the Open Data Value Paradox. *Proceedings of the 16th International Conference on Wirtschaftsinformatik*. <https://aisel.aisnet.org/wiz2021/HDigitaltransformation17/Track17/3>
- Enders, T., Benz, C., Schüritz, R., & Lujan, P. (2020). How to Implement an Open Data Strategy? Analyzing Organizational Change Processes to Enable Value Creation by Revealing Data. *Proceedings of the 28th European Conference on Information Systems*. 28th European Conference on Information Systems.
- EU Open Data Portal. (2022). EU Open Data Portal. <https://data.europa.eu/en>
- European Commission, Capgemini Consulting, Intrasoft International, Fraunhofer Fokus, con.terra, Sogeti, Open Data Institute, Time.lex, & University of Southampton. (2015). *Creating Value through Open Data Study on the Impact of Re-use of Public Data Resources*. Publications Office of the European Union.
- GLEIF. (2019). *GLEIF Registration Authorities List*. GLEIF. <https://www.gleif.org/en/about-lei/code-lists/gleif-registration-authorities-list>
- Global Open Data Index. (2015). *Company Register*. <https://index.okfn.org/dataset/companies/>

- Goldkuhl, G., Lind, M., & Seigerroth, U. (1998). Method Integration: The Need for a Learning Perspective. *IEE Proceedings - Software*, 145, 113.
- Gregor, S. (2006). The Nature of Theory in Information Systems. *MIS Quarterly*, 30(3), 611–642.
- Hendler, J. (2014). Data Integration for Heterogenous Datasets. *Big Data*, 2(4), 205–215.
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, 28(1), 75–105.
- Immonen, A., Palviainen, M., & Ovaska, E. (2014). Requirements of an Open Data Based Business Ecosystem. *IEEE Access*, 2, 88–103.
- Jaakkola, H., Mäkinen, T., & Eteläaho, A. (2014). Open Data: Opportunities and Challenges. *Proceedings of the 15th International Conference on Computer Systems and Technologies*, 25–39.
- Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, 29(4), 258–268.
- Krasikov, P., Legner, C., & Eurich, M. (2021). Sourcing the Right Open Data: A Design Science Research Approach for the Enterprise Context. In *The Next Wave of Sociotechnical Design* (Vol. 12807, pp. 313–327). Springer.
- Krasikov, P., Obrecht, T., Legner, C., & Eurich, M. (2020). Is Open Data Ready for Use by Enterprises? *Proceedings of the 9th International Conference on Data Science, Technology and Applications*, 109–120.
- Krasikov, P., Obrecht, T., Legner, C., & Eurich, M. (2021). Open Data in the Enterprise Context: Assessing Open Corporate Data's Readiness for Use. In S. Hammoudi, C. Quix, & J. Bernardino (Eds.), *Data Management Technologies and Applications* (Vol. 1446, pp. 80–100). Springer International Publishing.
- Máchová, R., & Lněnička, M. (2017). Evaluating the Quality of Open Data Portals on the National Level. *Journal of Theoretical and Applied Electronic Commerce Research*, 12, 21–41.
- Marmier, A., & Mettler, T. (2020). Different Shades of Perception: How Do Public Managers Comprehend the Re-use Potential of Open Government Data? *Proceedings of the 41st International Conference on Information Systems*.
- Martin, S., Foulonneau, M., Turki, S., & Ihadjadene, M. (2013). *Risk Analysis to Overcome Barriers to Open Data*. 11(1), 348–359.
- Masip-Bruin, X., Ren, G.-J., Serral-Gracia, R., & Yannuzzi, M. (2013). Unlocking the Value of Open Data with a Process-Based Information Platform. *2013 IEEE 15th Conference on Business Informatics*, 331–337.
- Nayak, A., Bozic, B., & Longo, L. (2021). (Linked) Data Quality Assessment: An Ontological Approach. *Proceedings of the 15th International Rule Challenge*.
- Neumaier, S., Umbrich, J., & Polleres, A. (2016). Automated Quality Assessment of Metadata across Open Data Portals. *Journal of Data and Information Quality*, 8(1), 1–29.
- Open Government Working Group. (2007). *The 8 Principles of Open Government Data*. <https://opengovdata.org/>
- Open Knowledge Foundation. (2005). *The Open Definition: Defining Open in Open Data, Open Content and Open Knowledge*. <https://opendefinition.org/>
- OpenDataBarometer & World Wide Web Foundation. (2020, October). *Open Data Barometer*. https://opendatabarometer.org/?_year=2017&indicator=ODB
- Opendatasoft. (2022). *A Comprehensive List of 2600+ Open Data Portals in the World*. Open Data Inception. <https://opendatainception.io/>
- Osagie, E., Waqar, M., Adebayo, S., Stasiewicz, A., Porwol, L., & Ojo, A. (2017). Usability Evaluation of an Open Data Platform. *Proceedings of the 18th Annual International Conference on Digital Government Research*, 495–504.
- Paulheim, H. (2016). Knowledge Graph Refinement. *Semantic Web*, 8(3), 489–508.
- Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, 24(3), 45–77.
- Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data Quality Assessment. *Communication of the ACM*, 45(4), 211–218.
- Prat, N., Comyn-Wattiau, I., & Akoka, J. (2015). A Taxonomy of Evaluation Methods for Information Systems Artifacts. *Journal of Management Information Systems*, 32(3), 229–267.
- Reiche, K. J., Höfig, E., & Schieferdecker, I. (2014). Assessment and Visualization of Metadata Quality for Open Government Data. *Conference for E-Democracy and Open Government*, 335–346.
- Ren, G.-J., & Glissmann, S. (2012). Identifying Information Assets for Open Data. *2012 IEEE 14th International Conference on Commerce and Enterprise Computing*, 94–100.
- Ruijter, E., Grimmikhuijsen, S., van den Berg, J., & Meijer, A. (2018). Open Data Work: Understanding Open Data Usage from a Practice Lens. *International Review of Administrative Sciences*, 86(1), 3–19.
- Sandkuhl, K., & Seigerroth, U. (2019). Method Engineering in Information Systems Analysis and Design. *Software & Systems Modeling*, 18(3), 1833–1857.
- Schatsky, D., Camhi, J., & Muraskin, C. (2019). *Data Ecosystems: How Third-Party Information Can Enhance Data Analytics*. Deloitte. https://www2.deloitte.com/content/dam/insights/us/articles/4603_Data-ecosystems/DI_Data-ecosystems.pdf
- Sein, M. K., Henfridsson, O., Puroo, S., Rossi, M., & Lindgren, R. (2011). Action Design Research. *MIS Quarterly*, 35(1), 37–56.
- Strand, M., & Syberfeldt, A. (2020). Using External Data in a BI Solution to Optimise Waste Management. *Journal of Decision Systems*, 29(1), 53–68.

- Stróżyńska, M., Eiden, G., Abramowicz, W., Filipiak, D., Małyszko, J., & Węcel, K. (2018). A Framework for the Quality-based Selection and Retrieval of Open Data. *Electronic Markets*, 28(2), 219–233.
- van Hesteren, D., van Knippenberg, L., Weyzen, R., Huyer, E., & Cecconi, G. (2022). *Open Data Maturity Report 2021*. Publications Office of the European Union. <https://data.europa.eu/doi/10.2830/394148>
- Van Nuffelen, B., Janev, V., Martin, M., Mijovic, V., & Tramp, S. (2014). Supporting the Linked Data Life Cycle Using an Integrated Tool Stack. In *Linked Open Data—Creating Knowledge Out of Interlinked Data: Results of the LOD2 Project* (Vol. 8661, pp. 108–129). Springer.
- Vetrò, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., & Morando, F. (2016). Open Data Quality Measurement Framework: Definition and Application to Open Government Data. *Government Information Quarterly*, 33(2), 325–337.
- Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, 12(4), 5–33.
- Welle Donker, F., & Van Loenen, B. (2017). How to Assess the Success of the Open Data Ecosystem? *International Journal of Digital Earth*, 10(3), 284–306.
- YData Labs. (2023). *Pandas profiling overview*. <https://pandas-profiling.ydata.ai/docs/master/index.html>
- Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality Assessment for Linked Data: A Survey. *Semantic Web*, 7(1), 63–93.
- Zhang, R., Indulska, M., & Sadiq, S. (2019). Discovering Data Quality Problems: The Case of Repurposed Data. *Business & Information Systems Engineering*, 61(5), 575–593.
- Zhang, S., Zhang, C., & Yang, Q. (2003). Data Preparation for Data Mining. *Applied Artificial Intelligence*, 17(5–6), 375–381.
- Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R., & Alibaks, R. S. (2012). Socio-Technical Impediments of Open Data. *Electronic Journal of E-Government*, 10(2), 156–172.
- Zuiderwijk, A., Janssen, M., & Janssen, M. (2014). Barriers and Development Directions for the Publication and Usage of Open Data. In *Open Government* (Vol. 4, pp. 115–135). Springer.
- Zuiderwijk, A., Janssen, M., Poulis, K., & van de Kaa, G. (2015). Open Data for Competitive Advantage. *Proceedings of the 16th Annual International Conference on Digital Government Research*, 79–88.

Essay 5

Introducing a Data Perspective to Sustainability: How Companies Develop Data Sourcing Practices for Sustainability Initiatives

Pavel Krasikov and Christine Legner

Faculty of Business and Economics (HEC), University of Lausanne, Switzerland

First version presented at the pre-ICIS Special interest Group on Green IS (SIGGreen) workshop,

2022

Extended version accepted to the special issue of the Communications of the Association for Information Systems (CAIS) on Digital Innovation for Social Development and Environmental

Action, 2023

Abstract: *Many companies use the UN Sustainable Development Goals as a point of reference for their sustainability initiatives and actions. Reporting on these goals requires collecting, processing, and interpreting substantial amounts of data (e.g., on emissions or recycled materials) that were previously neither captured nor analyzed. Although prior studies have occasionally highlighted the issues of data availability, data access, and data quality, a research void prevails on the data perspective in the sustainability context. This article aims at developing this perspective by shedding light on data sourcing practices for the reliable reporting of sustainability initiatives and goals. We make a two-fold contribution to sustainability and Green IS research: First, as theoretical contribution, we propose a framework based on institutional theory to explain how companies develop their data sourcing practices in response to regulatory, normative, and cultural-cognitive pressures. Second, our empirical contributions include insights into five case studies that represent key initiatives in the field of environmental sustainability, that touch on first, understanding the ecological footprint, and second, obtaining labels or complying with regulations, both on product and packaging levels. Based on five case studies, we identify three data sourcing practices: sense-making, data collection, and data reconciliation. Thereby, our research lays the foundation for an academic conceptualization of data sourcing in the context of sustainability.*

Keywords: Data sourcing, Sustainability reporting, Institutional theory, Data quality, Sustainable development goals, Triple bottom line

Table of contents

1	Introduction	166
2	Related work.....	169
2.1	Sustainability from the perspective of institutional theory	169
2.2	SDG implementation and sustainability reporting.....	170
2.3	The missing data perspective in Green IS.....	172
3	Methodology	175
3.1	Exploratory phase	176
3.2	Case selection.....	176
3.3	Data collection	178
3.4	Within- and cross-case analyses	179
4	Institutional pressures, sustainability initiatives, and data sourcing	182
4.1	Research framework.....	182
4.2	Sustainability initiatives and data sourcing practices on product level	184
4.3	Sustainability initiatives and data sourcing practices on packaging level	187
5	Data sourcing practices for sustainability.....	190
6	Discussion, implications, and future action	193
6.1	Summary of contributions and discussion	193
6.2	Implications for research	194
6.3	Implications for practice	196
6.4	Limitations and outlook	196
7	References.....	198
	Appendix 1	201
	Appendix 2.....	202

List of figures

Figure 15. Research framework.....	182
Figure 16. Analyze the ecological footprint on the product level (Type 1a)	186
Figure 17. Obtain a product label (Type 2a).....	187
Figure 18: Analyze the ecological footprint on the packaging level (Type 1b).....	188
Figure 19. Comply with packaging regulations (Type 2b)	189
Figure 20: Data requirements for sustainability initiatives at product and packaging level	192

List of tables

Table 37. Data-related problems in Green IS literature	173
Table 38. Research process	175
Table 39. Companies involved in the research process.....	176
Table 40. In-depth case studies (focus on product and packaging)	179
Table 41. Sustainability initiatives and data sourcing practices	184
Table 42. Data sourcing practices for sustainability.....	190
Table 43. Interviewee profiles	201
Table 44. List of definitions of the attributes	202

1 Introduction

The year 2015 marked the appearance of the United Nations (UN) General Assembly's agenda to ensure a more sustainable future by 2030 through Sustainable Development Goals (SDGs): a collection of 17 interlinked objectives that emphasize the interconnected environmental, social, and economic aspects of sustainable development (United Nations, 2022). Since then, sustainability objectives have become a priority of public and private sectors worldwide. The SDGs are widely used as a reference point, even though their operationalization and implementation at the local level – referred to as “localization” (Tremblay et al., 2021) – remains challenging. Many organizations work on mapping the SDGs into their own initiatives and actions (Corbett & Mellouli, 2017; Pan et al., 2022; Tremblay et al., 2021) and integrating them into their annual reports, emphasizing the importance of a holistic view on economic profit, as well as social and environmental impact, also known as the ‘triple bottom line’ (Milne & Gray, 2013). In general, sustainability reporting has significantly evolved over the past decades. Whereas standard formats for sustainability reporting lacked in the past (Melville, 2010), much progress has since been made due to the introduction of mandatory sustainability regulations (Christensen et al., 2021) within the frameworks of, among others, the Global Reporting Initiative (GRI) (GRI, 2022), the European Union (EU)'s (2014) Non-Financial Reporting Directive (NFRD) and its recently published expansion with the Corporate Sustainability Reporting Directive (CSRD) (Official Journal of the European Union, 2022), and the European Green Deal (European Commission, 2022).

Although the sustainability reporting structure and requirements are much clearer at present, data availability, data access and data quality have emerged as the main issues (Deloitte, 2021; EDM Council, 2022; Stoll, 2022). In reality, reporting on sustainability goals requires collecting, processing, and interpreting large amounts of data, especially on emissions and product composition, which have not been systematically collected or analyzed previously. Even when organizations can gather the required data, they often have to rely on estimates, and also lack details about its provenance. This drawback does not only compromise the reliability of the calculated sustainability indicators, but it also raises concerns about greenwashing (Szabo & Webster, 2021). For instance, the European Commission (2021) contends that 42% of analyzed green claims were “exaggerated, false or deceptive” with 59% of them lacking supportive evidence, while the United Kingdom's Competition and Markets Authority (2021) states that “40% of green claims could be misleading.” Without commonly accepted definitions and standards, it is difficult to collect and compare data on sustainability initiatives within and

across organizations; an aspect that has only been discussed within the scope of larger public initiatives, for example, in agriculture (Vrolijk et al., 2016) and in the European open data plan for the collection of geospatial, earth observation, or mobility data (Nuthi, 2022), but remains to be addressed in the enterprise context.

Despite the data's relevance to reliably report on sustainability initiatives and goals, there is a void of research on the data perspective in the context of sustainability. This also applies to Green IS (Pan et al., 2022; Seidel et al., 2017; Watson et al., 2010) and more specifically to environmental management information systems (EMIS), which are supposed to play a significant role in "structured and goal-oriented data gathering, administration, integration, and processing" (Stindt et al., 2014, p. 2) of environmental information. Although certain authors highlight data availability and data quality as key issues in EMIS (Melville et al., 2017; Zampou et al., 2022), the existing studies focus on EMIS design, in terms of components, features and design principles. Although they mention data quality as key concern, they hardly elaborate on the data requirements for EMIS and only give minimal attention to data accessibility for sustainable development (Machado Ribeiro et al., 2022). These data-related problems become particularly urgent when reporting on sustainability initiatives becomes mandatory and requires the audit of the reported information, as imposed under the CSRD (Official Journal of the European Union, 2022). To fulfill these requirements and ensure trustworthy sustainability reporting, companies need to build processes and practices to collect reliable, high-quality data not only within their own premises, but also externally, for instance, from suppliers.

Our study is a first step toward the development of a data perspective on sustainability and draws the attention to data sourcing, which is defined as "...procuring, licensing, and accessing data (e.g., an ongoing service or one-off project) from an internal or external entity (supplier)" (Jarvenpaa & Markus, 2020, p. 65). Institutional theory has been widely used in management and sustainability literature (Butler, 2011; Glover et al., 2014; Wang et al., 2015) to study the management practices that enterprises have developed to address regulative, normative, and cultural-cognitive pressures in their environment. Thus, assuming that pressures have an undeniable impact within the sustainability context (Daddi et al., 2020), institutional theory provides a useful lens to study how exogenous factors from the environment shape the emerging data sourcing practices in sustainability initiatives, and to address the following research question:

How do companies develop data sourcing practices in response to institutional pressures in the sustainability context?

In our study, we leverage qualitative research methods, including focus groups and case studies, which are “well-suited to capturing the knowledge of practitioners and developing theories from it” (Benbasat et al., 1987, p. 370). Our research setting is a multiyear research program studying data management practices for sustainability, which provides us with privileged access to data experts representing 12 multinational companies. These multinational firms are experiencing a wide range of institutional pressures and, in turn, report on their sustainability initiatives as part of their annual statements or in special corporate sustainability reports. For the case analysis, we collected data on key sustainability initiatives in the field of environmental sustainability in five manufacturing firms and analyzed them through the prism of institutional theory. This approach enables us to identify causal chains leading from the relevant pressures to the resulting sustainability initiatives, and to identify emerging data sourcing practices.

Our findings are a first step towards the development of a data perspective to sustainability and Green IS research. They make a two-fold contribution to sustainability and Green IS research. First, as theoretical contribution, we propose a framework based on institutional theory, to explain how companies – in the sustainability context – develop their data sourcing practices in response to exerted pressures. Second, our empirical contributions include insights into five case studies that represent key initiatives in the field of environmental sustainability, that touch on first, understanding the ecological footprint, and second, obtaining labels or complying with regulations, both on product and packaging levels. We derive three general data sourcing practices – sense-making, data collection, and reconciliation – which pave the way towards reliable and trustworthy sustainability reporting. Our study exemplifies impact-oriented Green IS research (Gholami et al., 2016), guiding enterprises on their way to become more sustainable by embedding sustainability in IS and in practice (Seidel et al., 2017).

The remainder of the paper is structured as follows: Section 2 introduces institutional theory to study how companies adapt their management practices in the context of sustainability. By reviewing prior literature related to SDGs, sustainability reporting and Green IS, this section identifies and justifies the missing data (sourcing) perspective as a research gap. Section 3 elaborates on our qualitative research design and the three phases of the research process. Section 4 synthesizes our research framework and the collected insights about product- and packaging-related initiatives, while section 5 generalizes our findings in the form of the three categories of data sourcing practices for sustainability. In section 6, we discuss our findings, derive implications for research and practice, and outline the limitations and actions to address them.

2 Related work

2.1 Sustainability from the perspective of institutional theory

The 2030 Agenda for Sustainable Development, adopted by all UN Member States in 2015, has made sustainability a high-priority, strategic topic of most organizations. At its heart are the 17 SDGs “which are an urgent call for action by all countries – developed and developing – in a global partnership” (United Nations, 2022), with overarching objectives to end poverty, improve health and education, reduce inequalities, and tackle climate change. Since 2015, enterprises have adopted the SDGs as a holistic framework to organize their own activities and clearly communicate their actions to the general public (Galleli et al., 2021). Prior research has shown that “institutional pressures influence organizations to address the Sustainable Development Goals” (Galleli et al., 2021, p. 5). These institutional pressures come from the environment and, more specifically, from the enterprises’ customers, their competitors, and the increasing number of regulations that impact on the prioritization of SDGs and sustainability initiatives (Galleli et al., 2021; Lu et al., 2018; Yang, 2018). According to institutional theory, which has been widely used in management and sustainability literature (Butler, 2011; Glover et al., 2014; Wang et al., 2015), organizations adapt their practices in response to a range of regulative, normative, and cultural-cognitive pressures in the environment. The theory argues that the resulting pressures, as perceived, incite enterprises to conform to institutional expectations (DiMaggio & Powell, 1983) since any violation thereof could jeopardize organizational performance and long-term development (Teo et al., 2003). DiMaggio & Powell (1983) discuss three types of institutional pressures (i.e., coercive, normative, and mimetic) that delimit and shape organizational actions. Building on the work of the early institutionalists, recognizing the multidisciplinary nature of the field, and connecting theory with empirical research, Scott (2013) conceptualizes “three pillars” that encapsulate regulative, normative, and cognitive pressures, which respectively relate to legally mandated behavior, behavior guided by moral norms, and commonly understood behavior. Organizations in a particular industry tend to adopt comparable practices and structures to establish their place and gain legitimacy within this industry (Scott, 2013).

Applied to sustainability, **regulative** pressures (also referred to as coercive or legislative pressures) originate from political influence and governmental agencies and result in legally imposed rules, laws, or sanctions. In this regard, sustainability regulations are among the major sources of external influences that drive environmental management practices (Butler, 2011;

Yang, 2018). They are typically delivered to enterprises in the form of environmental conventions/directives, for example, the EU's NFRD (Official Journal of the European Union, 2014) that all 28 EU members have adopted and incorporated in their respective national law. **Normative** pressures imply that companies go beyond the legal requirements and adopt new practices which conform with societal norms and values (Scott, 2013). In the context of sustainability, this is reflected in pressures exerted by customers. For instance, consumers' awareness of the ecologic, social, and economic consequences of their consumption drives the increase in their demands for more sustainable products (Lu et al., 2018). Enterprises, in turn, react to the changing demand by improving their supply chain practices (Lu et al., 2018; Yang, 2018). Finally, **cultural-cognitive** pressures (also known as mimetic pressures) are mainly driven by uncertainty and enterprises' ambiguity when stimulated by the environment (Scott, 2013). From the sustainability perspective, competitors' actions create precedents that prompt other enterprises to improve and mimic their environmental activities, among others, by reducing pollution and building a corporate green image (Yang, 2018). Another instance of a cultural-cognitive influence on enterprises is exemplified by The Carbon Disclosure Project, which motivates organizations to voluntarily evaluate and disclose their carbon dioxide emissions as well as their mitigation strategies to reduce the effects of climate change and to improve their corporate image (Butler, 2011; Melville, 2010).

These three types of institutional pressures and their influence on management practices are studied in both sustainability and Green IS research. For instance, Raj et al (2020) identify and discuss public procurement practices in different business contexts (e.g., sustainable supply chain, product modularity, environmental innovation). Butler (2011) uses the foundations of institutional theory to reflect on the implementation of a specific Green IS, the Compliance-to-Product application, as well as on its ability to support sense-making, decision-making, and knowledge sharing or knowledge creation. To conclude, institutional theory provides a widely accepted framework to study how enterprises adapt management practices, including the practices that drive SDG implementation and sustainability reporting.

2.2 SDG implementation and sustainability reporting

Although enterprises have made major efforts to address the 17 SDGs, reporting on these efforts is not without its challenges, and practitioner reports highlight data quality among the key concerns (Deloitte, 2021; EDM Council, 2022; Stoll, 2022). Most large enterprises do use the SDGs as a guiding framework to build a compelling narrative that conveys their achievements through corporate sustainability reports, but struggle to actually report on the SDGs, with any

real clarity. Despite the existence of SDG targets, which elaborate on the specific objectives of overarching SDGs, they face difficulties in operationalizing and mapping them to their own initiatives (Corbett & Mellouli, 2017; Pan et al., 2022; Tremblay et al., 2021). An interesting approach was adopted by Bissinger et al. (2020) who analyzed 232 voluntary sustainability standards (VSS) with more than 800 requirements, mapping them – in the process – to 16 (of the 17) SDGs. This empirical study was one of the first attempts to better understand how diverse standards contribute to the SDGs. However, the authors' findings revealed that a single standard could span multiple SDGs and could induce multiple overlaps, thus pointing toward a clear need to develop suitable KPIs and frameworks, developing synergies between the goals of the VSS and the UN.

While literature on SDG implementation and reporting remains scarce, sustainability reporting in general has been widely discussed. Originating largely from the triple bottom line framework (Milne & Gray, 2013), the aim of sustainability reporting is to obtain, process, and disseminate information (qualitative and quantitative) about the impact of enterprises on the economy, the environment, and people (Marx Gómez & Teuteberg, 2015; Seethamraju & Frost, 2016). Sustainability reporting has also evolved over the past decades, transcending the traditional financial reporting (Sisaye, 2021), particularly under the influence of new regulations and clearer sustainable development goals. Recent developments in the domain of integrated reporting show tangible progress in terms of reporting regulations (Christensen et al., 2021), particularly with reference to the GRI (GRI, 2022), the EU's NFRD (Official Journal of the European Union, 2014), and the European Green Deal (European Commission, 2022). The GRI provides clear guidelines on integrated reporting and much-needed metrics, and has been adopted on a large scale. Since its introduction, more than 10,000 companies – covering more than 100 countries and including 73% of the world's 250 largest firms – voluntarily chose the GRI (Christensen et al., 2021; GRI, 2022).

As sustainability and financial reporting are intrinsically linked (Sisaye, 2021), sustainability reporting initiatives within organizations are oftentimes driven by accounting and finance departments. Emerging in the context of traditional reporting (Sisaye, 2021), sustainability reporting stemmed as a standalone type with direct implications for reporting service providers (e.g., consulting and audit firms). "Triple-bottom-line reporting, also known as sustainability reporting, involves reporting nonfinancial and financial information to a broader set of stakeholders than just the shareholders" (Ivan, 2009, p. 108). In addition, environmental and social concerns of Corporate Social Reporting (oftentimes directly associated with sustainability reporting) go hand in hand with financial reporting and are associated with competitive edge

and improvements in financial performance (Sisaye, 2021). Furthermore, reporting is largely led by advisory institutions (e.g., the World Resources Institute and the United Nations Conference on Trade and Development) or by professional accountancy bodies (e.g., the Federation of Accountants, the Federation of European Accountants, Deloitte, KPMG, PWC and Ernst & Young) (Seethamraju & Frost, 2016). Hence, financial reporting integrates sustainability information to identify financial risks or opportunities related to the impact of the reporting entity's activities and, in turn, reports on enterprises' assets, liabilities, equity, and expenses.

2.3 The missing data perspective in Green IS

Sustainability initiatives and the adjoining reporting rely on qualitative and quantitative information about the companies' actions. To this end, "some organizations have sophisticated information systems that are capable of collecting, storing and analyzing certain types of sustainability information" (Frost et al., 2012, p. 224). This has also motivated researchers to study EMIS (Bansal & Roth, 2000; El-Gayar & Fritz, 2006; Walls et al., 2011), as a subfield of Green IS, which are "organizational-technical systems for systematically obtaining, processing, and making available relevant environmental information available in companies" (El-Gayar & Fritz, 2006, p. 768). Although EMIS are seen as enablers "for structured and goal-oriented data gathering, administration, integration, and processing" (Stindt et al., 2014, p. 2), most of the studies focus on system design and adoption rather than on the data as such. These studies include prototypes and investigations into EMIS implementation (Teuteberg & Straßenburg, 2009), as well as design principles for developing sustainable reporting or monitoring systems for emissions and energy usage (Hilpert et al., 2014; Zampou et al., 2022). They mainly focus on EMIS components and functional features, such as supply chain coordination or reporting (Zampou et al., 2022), as well as information flows to combine various sources and calculate KPIs. Although Zampou et al. (2022) highlight the importance of considering data quality in the EMIS design, such as data-cleansing process (e.g., to estimate missing product weights or volumes of product categories), existing EMIS literature has not further elaborated on data-related requirements or processes.

This reflects the current state of Green IS literature, which consistently reports that data-related problems are among the primary challenges that practitioners and researchers face when dealing with sustainability data (Marx Gómez & Teuteberg, 2015; Melville et al., 2017; Watson et al., 2010; Zampou et al., 2022). Several authors criticize the accessibility of data for sustainable development (Machado Ribeiro et al., 2022) and the unattended challenge "to gather all required sustainability data from external and internal sources" (Seethamraju & Frost, 2016, p.

3). The main data-related problems, identified in Green IS literature (Table 37), include the unavailability of data or simply unknown data (Machado Ribeiro et al., 2022; Watson et al., 2010; Zampou et al., 2022), the lack of data integration and consolidation (Marx Gómez & Teuteberg, 2015; Zampou et al., 2022), and insufficient attention given to data quality and the underlying dimensions thereof, namely completeness and accuracy (Machado Ribeiro et al., 2022; Melville et al., 2017; Zampou et al., 2022). However, the existing studies do not go beyond stating the data-related problems nor do they elaborate on the specific data requirements or practices to address the issues.

Source	Context	Problem statements	Problem category
Watson et al. (2010)	Development of energy information systems targeted at analyzing and reducing energy consumption	“...the major issue is to design a sensor network that provides sufficient granularity to provide adequate data for an optimal solution” (Watson et al., 2010, p. 29) “what data should be reported by an energy information system to inform governments' energy policies?” (Watson et al., 2010, p. 30)	Granularity of data Unknown data
Marx Gómez & Teuteberg (2015)	Development and technical features of Corporate EMIS (CEMIS)	Lack of integration measures and a consolidation of various sources in an enterprise setting	Data integration and consolidation
Melville et al. (2017)	Systems that enable organizations to adopt low-carbon operations	Typical data quality dimensions, such as completeness and accuracy, are not sufficiently addressed	Data quality
Machado Ribeiro et al. (2022)	Literature review on data governance and sustainability	Importance of data governance mechanisms for sustainability, pointing toward the importance data accessibility and data quality aspects	Data accessibility and data quality
Zampou et al. (2022)	Design theory for Energy and Carbon Management Systems	“...challenges in terms of, for example, data quality and availability, data capturing and integration, and information sharing” (Zampou et al., 2022, p. 6)	Data quality and availability, data capturing and integration, and information sharing

Table 37. Data-related problems in Green IS literature

As noted earlier, data is undeniably important to reliably report on sustainability initiatives and goals along the existing frameworks. Thus, data needs to be an integral part of enterprise sustainability activities, and more research is needed on the data perspective in the context of sustainability. To address the identified data-related problems, companies must understand the data requirements and develop dedicated data sourcing practices; unfortunately, neither sustainability nor Green IS research has embraced these topics. For instance, if data sources' granularity is inadequate or is not consolidated before integration, the resulting data may be

incomplete, inconsistent, or inaccurate, which can negatively impact the decisions based on this data. The increasing number of required data sources and their heterogeneity also call for the development of enterprise-wide data sourcing practices rather than ad-hoc sourcing.

3 Methodology

In view of our research goals, the present study employs institutional theory to explore the relevant pressures shaping sustainability initiatives and the subsequent organizational responses in the form of data sourcing practices. We leverage qualitative research methods, including focus groups and case studies, which are well suited to grasp the richness of specific situations in naturalistic settings (Benbasat et al., 1987; Van de Ven & Poole, 2005). Our research setting provided us with privileged access to data experts representing 12 multinational companies that have made SDG-related commitments and were in the process of refining their data sourcing practices. All companies are large multinational companies (or incumbents) from highly institutionalized industries (Powell & DiMaggio, 2012); they are characterized by a high level of regulation, standardization, and formalization that require their conformity and adherence to pressures. We closely collaborated with these companies in a multiyear research program studying data management practices for sustainability, subdivided into three research phases (see Table 38).

Research phases	Phase 1: Exploratory research (05/2021 – 02/2022)	Phase 2: Five case studies (02/2022 – 04/2022)	Phase 3: Within- and cross-case analysis (04/2022 – 09/2022)
Objectives	<ul style="list-style-type: none"> • Explore sustainability reporting and the data-related challenges • Identify the most relevant and tangible sustainability initiatives in the participating companies 	<ul style="list-style-type: none"> • Gain in-depth understanding of the ongoing sustainability initiatives in five selected companies • Obtain insights into the data sourcing requirements, challenges, and emerging practices 	<ul style="list-style-type: none"> • Analyze institutional pressures which influence the sustainability initiatives, and the resulting data sourcing practices within and across the case studies • Generalize and validate the results with experts
Activities	<ul style="list-style-type: none"> • Focus group 1 (5 participants from 5 companies): data challenges in the sustainability reporting process • Focus group 2 (8 participants from 8 companies): scoping of the reporting goals • Focus group 3 (17 participants from 12 companies): identification of sustainability initiatives 	<ul style="list-style-type: none"> • Primary data: 60-minute individual, semi-structured interviews with 5 experts from 5 companies • Secondary data: internal company documentation and presentations, corporate sustainability reports • Prepare a write-up per case, comprising key statements and a process map of the data sourcing activities, and validate it with the experts 	<ul style="list-style-type: none"> • Within-case analysis: coding of each case as standalone entity based on framework that builds on institutional theory • Cross-case analysis: search for patterns across cases • Focus group 4 (5 participants from 4 companies): consolidation of data sourcing practices • Focus group 5 (10 participants from 8 companies): data model for data sourcing
Outcomes	<ul style="list-style-type: none"> • Problem scoping, list of 12 sustainability initiatives 	<ul style="list-style-type: none"> • Five case write-ups and process maps 	<ul style="list-style-type: none"> • Framework building on institutional theory, three sourcing practices

Table 38. Research process

3.1 Exploratory phase

Our research activities began with an exploratory phase during the period of May 2021 to February 2022. We started with two focus groups involving 13 data management experts from 13 multinational companies with the aim of understanding the status of and issues in their sustainability reporting. This enabled us to identify key problem areas, among others, the ambiguous data requirements, ad-hoc data sourcing practices, and accompanying data quality-related issues. To narrow our scope, we subsequently conducted a third focus group with representatives of 12 multinational companies with the goal of identifying ongoing and concrete sustainability initiatives among the group (see Table 39). Although sustainability reporting remains an overarching driver, we found that many companies had also defined key sustainability initiatives and developed dedicated data sourcing practices to address them.

Company	Industry	Revenue/employees	Key informants	Key sustainability initiatives
A*	Fashion and retail	\$1B–50B/~60,000	Director data governance	Product labeling
B*	Engineering and electronics	\$1B–50B/~400,000	Director master data management	Product ecological footprint
C*	Pharmaceutical, chemicals	\$1B–\$50B/~100 000	Head of product data management	Product labeling
D*	Manufacturing, chemicals	\$1B–\$50B/~5,000	Data steward material & product	Plastic packaging tax
E*	Consumer goods	\$50B–\$100B/~350,000	Global master data lead	Packaging recyclability
F	Adhesive & beauty products manufacturing	\$1B–\$50B/~20 000	Director master data	Consumption of material in packaging
G	Manufacturing, chemicals	\$1B–\$50B/~20 000	Head of data management	Sustainability reporting
H	Logistics	\$1B–\$50B/~70'000	Program manager governance	ESG reporting, emission along the supply chain
I	Software development	\$1B–\$50B/~100 000	Solution advisor expert	Reduction of workplace inequalities
H	Manufacturing, automotive	\$1B–\$50B/~90 000	Senior data and analytics governance professional	Centralized sustainability reporting
J	Packaging, food processing	\$1B–\$50B/~25 000	Enterprise data governance manager	Circular business models, advanced analytics
K	Manufacturing, automotive	\$1B–\$50B/~150 000	Senior data architect	Supply chain emissions

*All companies were involved in Phases 1 and 3, * indicates the companies involved in Phase 2*

Table 39. Companies involved in the research process

3.2 Case selection

From February to April 2022, we immersed ourselves in the data sourcing practices of five of the 12 companies (see Table 40), thereby contributing to in-depth case studies. According to Benbasat et al. (1987), case studies are well suited to capture practitioners' knowledge and develop theories based thereon. Multiple case studies improve external validity while supporting analytical generalization (Yin, 2009). Although all 12 represent large, product-oriented, multinational companies from highly institutionalized industries that currently focus

on sustainability goals and commitments, they had reached different levels of maturity in their data sourcing practices and ongoing sustainability initiatives. Using purposeful sampling, we selected the five most mature companies (of the 12) for further investigation. This maturity was reflected by the progress made in their sustainability initiatives and the supporting evidence for a systematic approach to sustainability reporting. Additionally, by selecting five companies representing different industries and positions in the value chain, we expected natural variation with regard to sustainability initiatives and related data sourcing practices, and to better determine the influence of environmental pressures.

Being active in the fashion and retail industry, Company *A* faces an increasing awareness of sustainability and fixed aggressive goals to increase the use of recycled materials. In their annual reports, *A* announced their commitment to end plastic waste, backed with strong objectives to reduce greenhouse gases emissions, use of sustainable materials for their products, and ultimately achieving climate neutrality across the whole value chain. It introduced product labels (e.g., 100% recycled polyester) to better communicate its progress to consumers, but faced challenges in collecting the relevant information due to a high level of outsourcing to third-party suppliers in Asia. Company *B*, representing the engineering and electronics industry, is actively engaged in improving the traceability of product-related emissions. With their objective to be honest and transparent in their sustainability reporting, *B* strives for comprehensible benchmarks and understandable metrics. To systematically collect and manage data on the ecological footprint of its materials and components, *B* is currently redefining the structure of its product master data in ERP systems. Company *C*, in line with pharmaceutical and chemical industry requirements, engages in the transparent communication of product composition, particularly in terms of specific substances. Certain substances like heavy metals, which are necessary for chemical synthesis processes, demand careful consideration and pre-treatment to ensure they are properly managed before being discharged into wastewater. Being committed to sustainable use of resources and respecting planetary boundaries, *C* aims at obtaining certifications (e.g., Wildlife Habitat Council certification) and transparently communicating product composition through customer-facing third-party certification labels. Company *D*, operating in the manufacturing and chemicals industry, embraced their path on becoming more sustainable, including reduction of emissions, responsible sourcing, use of recycled and bio-based materials, circularity, and reduction of waste. Among all, *D* also faces new regulations (e.g., the UK plastic packaging tax) which impose penalties if the specified quantities of recycled plastic contained in packaging are not met. Thus, *D*'s main objective is to comply with emerging international regulations regarding packaging composition. Finally, company *E*, as one of the

global players in the consumer goods industry, has set ambitious objectives in terms of waste reduction and protection of nature, namely with significant reduction of plastic pollution. *E* attempts to meet increasing customer expectations by reducing the use of virgin plastic in different types of packaging. Consequently, *E* analyses the packaging's composition to improve its recyclability, specifically regarding its combined components.

3.3 Data collection

We collected primary data by conducting semi-structured interviews with key informants – respectively representing each of the five companies (see Table 43 in the Appendix for the interviewee profiles) – between February and April 2022. We selected key informants who actively participate in the supervision and execution of data-related activities and the collection of data requirements in the ongoing sustainability initiatives. To observe incremental changes and capture the issues and challenges which accompany the implementation of sustainability initiatives, we ensured that the informants had significant tenure in their respective companies, as well as extensive experience in the field of data management. For each case, we conducted a one-hour interview per interviewee to garner insights about the company's sustainability initiatives, underlying data requirements, and emerging data sourcing practices. As an instrument of inquiry, our semi-structured interviews followed a nominal protocol that allowed us to ask questions related to the aims of the study (Castillo-Montoya, 2016) while simultaneously maintaining the fluidity and openness of the discussion with the interviewees. As part of the interviews, we jointly documented data sourcing activities together with the interviewees on a Miro board (collaborative digital whiteboard platform), starting from the core business processes, involved roles, required data objects, and encountered challenges. We then developed a first version of a process map for each company that included associated data objects and the relationships between them. After each interview, a write-up comprising key statements and the process maps for the documented initiatives were shared with the interviewee to confirm the correctness of the collected information and to clarify misunderstandings.

We complemented the interviews with an analysis of additional documents that we gathered throughout our research activities (e.g., slides shared during focus group presentations along with an overview of sustainability initiatives, the underlying process, involved applications, and data landscapes) and of publicly available information (e.g., corporate sustainability reports of the respective companies that detailed their goals and progress in achieving them). By combining primary and secondary sources, we triangulated the collected data to ensure

construct validity (Yin, 2009) and complemented the process maps with additional information about the company’s sustainability goals and context of the sustainability initiatives.

Company	Institutional pressures	Sustainability goal and related SDGs	Sustainability initiative	Data sourcing challenges
A	Cultural-cognitive: increased competition due to the appearance of more visible products using recycled materials Normative: increased customer demands for more sustainable products	Increase the use of recycled materials and better communicate the achieved progress to the end-consumers through self-declared product labels. SDGs: 3, 5, 6, 7, 8, 9, 10, 12, 13, 14, 17	Self-declared product labels	Inability to capture the data Insufficient data to perform the calculations Aggregating different units of measure
B	Normative: need for more professionalized approaches when communicating the ecological footprint of the products	Improve the traceability of the product-related emissions SDGs: 2, 3, 6, 7, 8, 9, 11, 12, 13	Product ecological footprint	Inability to capture the data Unclear roles and responsibilities in data sourcing processes Finding the right level of granularity for data aggregation
C	Regulative: legal requirements to clearly label the products based on the used substances Normative: moral obligations due to customer demands for clearer labeling	Obtain and clearly label the required product certifications SDGs: 1, 2, 3, 5, 6, 13, 15	Third-party product labels	Difficulties in aggregating the data Challenges faced when collecting the necessary data to comply with the certification requirements
D	Regulative: need to comply with plastic tax regulations in the concerned markets, e.g., UK plastic packaging tax	Comply with new regulations regarding the quantities of recycled plastic in the packaging SDGs: 8, 12, 13	Compliance with plastic packaging tax	Unclear how to deal with constantly changing regulations Identify what the regulations consider as packaging (avoid a possible confusion with the product itself)
E	Normative: increasing customer expectations regarding sustainable product packaging	Reduce the use of virgin plastic in the different types of packaging SDGs: 9, 12, 13, 14, 15, 17	Improve packaging recyclability	Lack of common definitions regarding what is considered recyclable Unclear how to deal with multiple packaging components, particularly when different elements are combines

Table 40. In-depth case studies (focus on product and packaging)

3.4 Within- and cross-case analyses

In the last phase of our research process, we analyzed the collected data, starting with the within-case analysis and then searching for patterns across the cases. For the within- and cross-case analyses, we used a research framework (see subsection 4.1), which we have developed by employing institutional theory to analyze and interpret our empirical insights.

We first investigated each case as a stand-alone entity and identified the causal links between institutional pressures and organizational responses in the form of sustainability initiatives and data sourcing practices. In line with our research objectives and to uncover the underlying conceptual logic of the collected case material (Miles et al., 2014), we analyzed each case based on the research model to identify for each case the relevant types of institutional pressures (regulative, cultural-cognitive, and normative), the prioritized sustainability initiatives, as well as data-related processes which companies go through in their reporting activities. As the process maps provide very rich data about the individual data sourcing practices, we inductively developed a coding scheme, i.e., using visual mapping (Miles et al., 2014), outlining the five key phases of the data sourcing process, namely planning, analyzing the relevant regulations and internal sustainability objectives, data collection and preparation, data integration, and finally reporting.

The within-case analysis provided a detailed understanding of the unique factors and context that influence the prioritization of sustainability initiatives, namely the motivations behind the engagement, documented in the activities of the planning phase within the process maps. These motivations were mapped with the institutional pressures from literature (see subsection 2.1), allowing to properly document the business context of the sustainability initiative, gathering insights into applicable regulations, and understanding the involved roles and their responsibilities. We carefully analyzed the process maps in order to understand how each company performed the remaining activities – starting from collecting data from internal and external sources (including suppliers or other third parties), defining the gaps and assessing the usability of the data, defining target architecture, integrating data and aggregating it for further manipulations and calculations. By conducting iterative coding, and maintaining construct validity through peer debriefings, the within-case analysis provided a rigorous examination of the data sourcing practices of each company in the context of sustainability initiatives.

After comprehending the dynamics of each case, we analyzed cross-case patterns to gradually build a rich conceptualization, creating types or groups to compare and examine cases for shared configurations (Miles et al., 2014; Yin, 2009). We employed pattern matching to identify recurring themes across the cases, namely in terms of the exerted pressures and types of the initiatives (see subsection 4.2.1 – 4.3.2), and the data sourcing practices. Based on the similarities and divergences in the five cases, we classified the initiatives along two dimensions, the scope (i.e., product and packaging) and goal of the sustainability initiative (i.e., analyzing the ecological footprint and complying with regulatory requirements or labels). We compared the five cases with regard to the activities across the process phases, and identified similarities as

well as the differences, such as involved stakeholders, involved data objects, and necessary KPIs for reporting. The cross-case analysis of the process maps allowed us to identify three data sourcing practices which emerged in all cases, namely sense-making, data collection, and data reconciliation (see section 5).

In a final step, we discussed our findings in two focus groups with the larger group of companies. We used the first of these focus groups to validate the three identified data sourcing practices and generalize their characteristics. In the second focus group, we discussed our insights into the specific data requirements for product and packaging levels, which we documented in a conceptual data model.

4 Institutional pressures, sustainability initiatives, and data sourcing

4.1 Research framework

To explain how data sourcing practices develop in the context of sustainability, we employ institutional theory as a theoretical lens (see subsection 2.1) and developed a research framework (see Figure 15), which defined the a-priori constructs to analyze the cases. On its left-hand side, the framework posits that the three types of institutional pressures – regulative, normative, and cultural-cognitive – influence organizations engaging in sustainability initiatives (Galleli et al., 2021). On the right-hand side, the framework comprises the data sourcing practices which enterprises develop in response to the exerted institutional pressures to support the reporting on the sustainability initiative. As data sourcing practices have not been previously studied, we inductively derived the three data sourcing practices - sense-making, data collection, and data reconciliation – from the within and cross-case analyses.

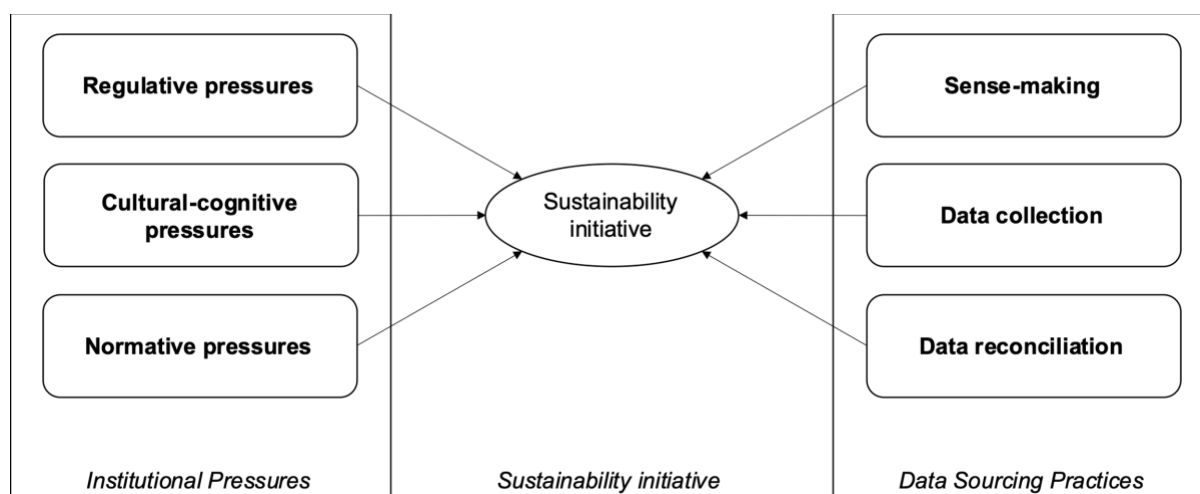


Figure 15. Research framework

In the five selected cases, we analyze two types of sustainability initiatives which are representative of manufacturing companies, and which can either apply to product or packaging level (see Table 41). The first type of initiative concerns the **ecological footprint** and requires an analysis of the materials that make up the product or its packaging. Based on the Bill of Material (BoM), it involves identifying, examining, and understanding the composition on the lowest level of granularity. The second type of initiative goes a step further and aims at **obtaining labels** (voluntary) or **complying with regulations** (mandatory). These initiatives

require an assessment of the material composition against the rules defined by regulations, product certification bodies (for a third-party label), or internally (for a self-defined label).

Based on the within- and cross-case analyses, we identify three categories of data sourcing practices that companies develop in their sustainability initiatives: sense-making, data collection, and data reconciliation. Due to the novelty of sustainability activities within the enterprises, **sense-making** involves a sophisticated analysis of the sustainability goals, ambitions, and regulations, as well as of their interpretation in terms of data requirements (Butler, 2011). During this phase, it is crucial to concretize and understand how to report on the sustainability initiatives to clarify the data requirements and identify the data objects and attributes which must be collected. It also requires a clarification of organizational matters, including the definition of roles and responsibilities to source the data. Based on these insights, **data collection** can be initiated to obtain the data needed for the sustainability initiatives. This data is often located within the existing operational systems (e.g., ERP or PLM systems), but it must be amended for the intended purpose of use. Due to value chain specialization and the specific data requirements thereof, some of the data has to be acquired externally (e.g., from suppliers or other parties). Finally, **data reconciliation** is necessary to prepare the data for further manipulations such as KPI calculations. This practice involves harmonizing the data obtained from different sources (with high variability of data types and formats) and aggregating it to the required level of granularity.

	Type 1: Analyze the ecological footprint	Type 2: Obtain the label or comply with the regulation
Product level	<p>Type 1a: analyze the consumption of critical materials at the product level (Cases A, B, C)</p> <p>Institutional pressures:</p> <ul style="list-style-type: none"> • Cultural-cognitive: increased competition due to the appearance of more visible products using recycled materials • Normative pressures: growing demand for sustainable products <p>Data sourcing practices:</p> <ul style="list-style-type: none"> • Sense-making: understand the product composition (BoM) and materials at the lowest level of granularity • Data collection: identify missing data for finished products (BoM) and related materials; collect them from internal and external sources • Data reconciliation: harmonize and standardize material classifications and descriptions 	<p>Type 2a: obtain customer-facing, self-declared product labels (Case A) or obtain a third-party product certification label (Case C)</p> <p>Institutional pressures:</p> <ul style="list-style-type: none"> • Regulative: labeling requirements for product components/substances • Cultural-cognitive: intensified competition arising from prominent products offerings • Normative: moral obligations due to customer demands for clearer labeling and more sustainable products <p>Data sourcing practices:</p> <ul style="list-style-type: none"> • Sense-making: understand the obtention conditions for the third-party or self-declared labels (e.g., the presence or absence of materials, thresholds) • Data collection: identify relevant data for label obtention within the finished product (BoM) and related materials • Data reconciliation: map material classifications and descriptions to label requirements; aggregate material data with different granularities to the product level
Packaging level	<p>Type 1b: analyze the recyclability of the materials used in packaging (Cases D, E)</p> <p>Institutional pressures:</p> <ul style="list-style-type: none"> • Normative: growing customer expectations for sustainable packaging. <p>Data sourcing practices:</p> <ul style="list-style-type: none"> • Sense-making: understand the packaging composition (BoM) at the lowest level of granularity • Data collection: identify missing data for packaging (BoM) and related material materials; collect them from internal and external sources • Data reconciliation: harmonize and standardize material classifications and descriptions 	<p>Type 2b: comply with the plastic packaging tax regulation (Cases D)</p> <p>Institutional pressures:</p> <ul style="list-style-type: none"> • Regulative: need to comply with plastic tax regulations in the concerned markets • Normative: growing demand for sustainable product packaging <p>Data sourcing practices:</p> <ul style="list-style-type: none"> • Sense-making: understand the limit set by plastic packaging tax (thresholds) • Data collection: identify relevant data for measurable thresholds and conditions within packaging (BoM) and related materials • Data reconciliation: map material classifications and descriptions and assess packaging composition against rules set by regulations

Table 41. Sustainability initiatives and data sourcing practices

4.2 Sustainability initiatives and data sourcing practices on product level

4.2.1 Type 1a: Analyze the ecological footprint

Cultural-cognitive and normative pressures drive product-related sustainability initiatives in companies A, B, and C that seek to analyze the ecological footprint (see Figure 16). Although tough competition and the appearance of more visible products using recycled materials have motivated these companies to reevaluate their product offerings, increased customer demands for more sustainable products have taken their toll on them, leading to their reduced

consumption of critical materials, such as plastics (to reduce their carbon footprint and to minimize environmental harm). To analyze the ecologic footprint, *A*, *B*, and *C* first had to examine and understand product composition at the lowest level of granularity (*sense-making*). While this seems obvious, the companies had to go beyond their usual manufacturing perspectives and report on their actual use of specific materials in their final products. For each final product, this implies investigating the as-is BOM that lists all the components and materials that went into the product, having been procured from suppliers or manufactured in the company's own plant.

First, to collect the necessary data, the companies assess and identify missing data for finished products and related materials and, second, collect the required data from internal and external sources. Interestingly, using available data to analyze the ecological footprint proved to be challenging, especially in the case of *A*. The reasons being a lack of required material classifications and that the product data was not previously analyzed within the ambit of the recycled materials used. In addition, data was often incomplete or not maintained within the enterprise due to increased levels of supplier outsourcing.

Once the data is collected, it must be harmonized and aggregated to calculate the percentage of specific materials at the level of the finished products. In this regard, companies do not only struggle to standardize material classifications and descriptions (e.g., external reference data GSI for chemical substances), but also have to aggregate them when using different units of measures such as weight and surface. Reconciliation primarily prepares the ground for further manipulations of product data and provides clarity about the final product's ecological footprint in terms of individually used materials.

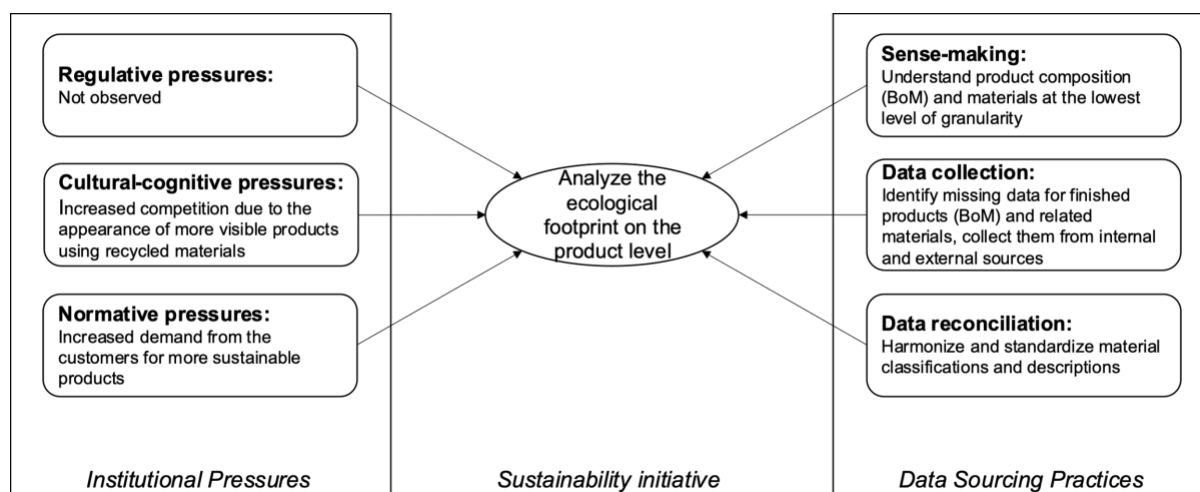


Figure 16. Analyze the ecological footprint on the product level (Type 1a)

4.2.2 Type 2a: Obtain a product label

The analysis of the ecological footprint is only the first step toward more ambitious sustainability initiatives with the aim of obtaining product certification labels (see Figure 17). As documented by cases A and C, all three institutional pressures – regulative, normative, and cultural-cognitive – impact the companies’ practices. First, there are legal requirements to clearly label certain products, based on the used substances (e.g., in the pharmaceutical industry). Second, positioning labeled products is an important distinguishing aspect that allows companies to differentiate themselves from their competitors and to secure price premiums. Third, companies have moral obligations toward their customers who demand clearer labeling and more sustainable products.

With this type of initiative, companies adapt their data sourcing practices by starting with sense-making of the obtention conditions for the third-party or self-declared labels. While there is no common way of specifying the conditions, they typically relate to the presence or absence of particular materials and define specific thresholds. Finding a suitable label also proves to be challenging since the obtention criteria require an interpretation of and alignment with the narratives that the companies intend to communicate about their products. In terms of data collection, they need to identify relevant data attributes for label obtention within the finished product (BoM) and related materials. Conclusively, for purposes of data harmonization, A and C map material classifications and descriptions to label requirements, and aggregate material data with different granularities to the product level. Since combinations lead to new requirements for a different product composition (e.g., in the pharmaceutical and chemical industry), it is noteworthy to consider the aggregations, which are done with multiple materials in a single product.

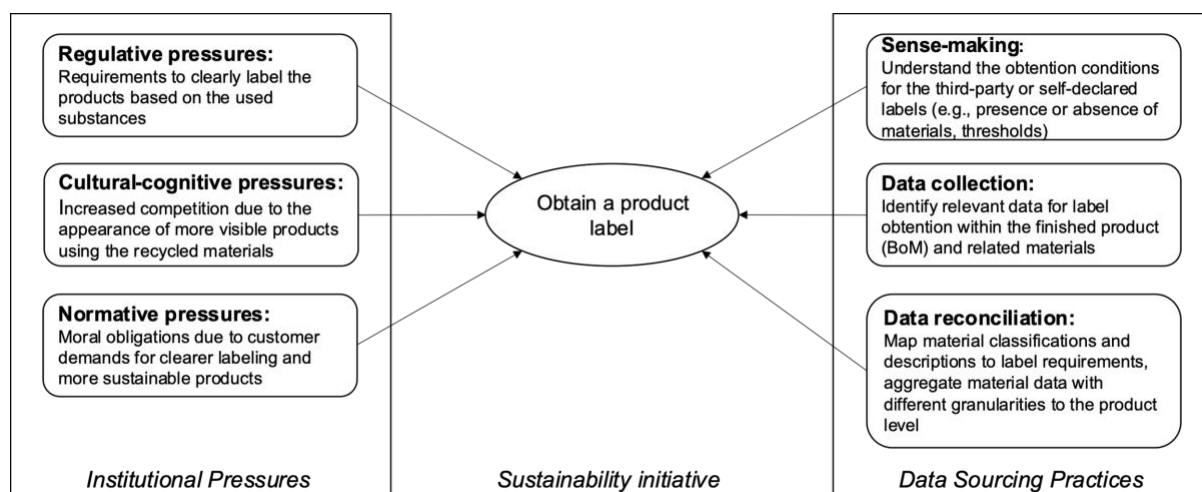


Figure 17. Obtain a product label (Type 2a)

4.3 Sustainability initiatives and data sourcing practices on packaging level

4.3.1 Type 1b: Analyze the ecological footprint

With the intention of assessing the recyclability of the used packaging materials, cases *D* and *E* analyze the ecological footprint at the packaging level (see Figure 18). We found that the regulative and cultural-cognitive pressures do not play a significant role in this initiative, although they increase the all-important customer expectations regarding sustainable product packaging (i.e., normative pressures).

In terms of sourcing practices, we found that – like the product level – understanding packaging composition (BoM) at the lowest level of granularity is not a trivial matter and requires sense-making. For instance, with different types of packaging, *D* and *E* convey the importance of setting a clear scope for the analysis, such as retail packaging, unit packaging, or packaging for protection during transportation. Relevant data must be collected to perform these analyses, starting with the identification of missing data for the packaging (BoM) and related materials which, in turn, is collected from internal and external sources. Material classifications and descriptions for the used packaging must be harmonized and standardized to perform the necessary calculations depicting the composition of the packaging. Company *E* highlights the importance and difficulties of the correct aggregation, since packaging types often tend to combine multiple components, some of which are entirely non-recyclable. *E*'s global master data lead states that: “a product can go through different states of packaging, from unit-level to pallet aggregation, which is a challenge from the data management perspective”. This

potentially leads to packaging confusion as a whole, especially when a material combination makes the entire packaging non-recyclable.

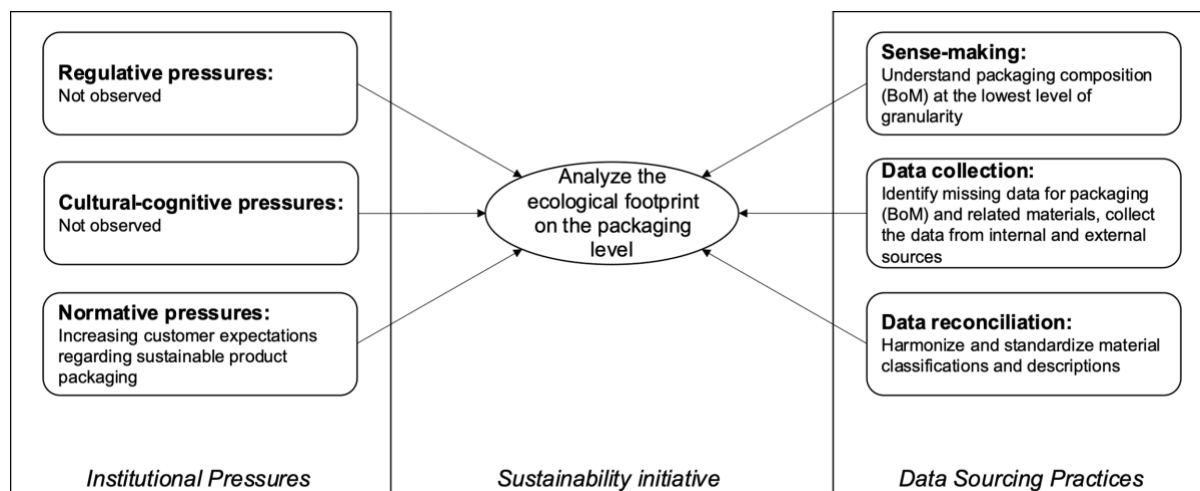


Figure 18: Analyze the ecological footprint on the packaging level (Type 1b)

4.3.2 Type 2b: Comply with regulations

Building on the obtained understanding of the packaging’s composition, companies can engage in additional initiatives. Case *D*’s objective is to comply with emerging regulations that specify acceptable thresholds of recycled plastic in the packaging (see Figure 19). Evidently, the regulative pressure and emerging regulatory requirements compel the companies to comply with and engage in such initiatives. These not only encompass country specific requirements concerning the consumption of manufactured single-use items (Italy) and the proportions of recycled plastic in a packaging component (UK), but also the companies’ own initiatives. Furthermore, similar to the previous initiatives, cultural-cognitive pressures in the form of customer expectations influence company practices. Therefore, *D* adopted a set of data sourcing practices. First, it is necessary to understand the limits set by the plastic packaging tax (e.g., in terms of the threshold for the presence of recycled plastics in the packaging) and by the incumbent tax rates. Consequently, it is necessary to distinguish between the packaging itself and the individual components which are used for the packaging (e.g., adhesive, liner, core, and backing). To illustrate, in the case of *E* this was of particular importance since the individual packaging components are also deemed to be products, thus complicating compliance and requiring an adaption of the components’ unit of analysis. Second, it is therefore necessary to identify relevant data for measurable thresholds and conditions within packaging (BoM) and related materials. Third, in terms of data reconciliation, material classification and descriptions must be mapped, along with an assessment of the product’s packaging composition against the rules set by the regulations.

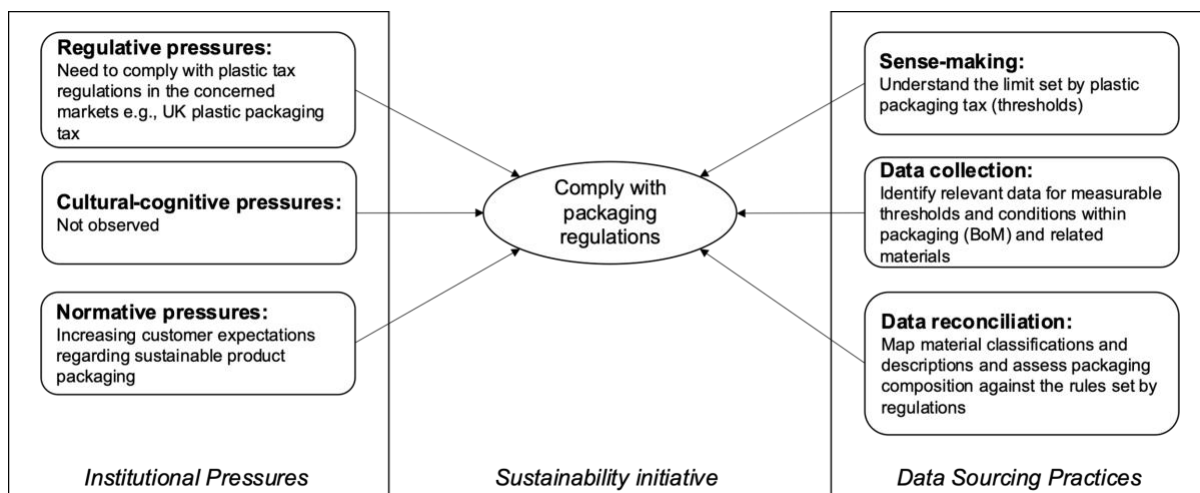


Figure 19. Comply with packaging regulations (Type 2b)

5 Data sourcing practices for sustainability

Based on our empirical findings, we identify three categories of data sourcing practices that we discuss in more detail with reference to their activities, outcomes, roles and responsibilities, and associated challenges (see Table 42).

Data sourcing practices	Sense-making	Data collection	Data reconciliation
Activities	Analyze and interpret the sustainability goals and identify the relevant data objects for and attributes of sustainability initiatives Decide on the approach to data collection and processing	Analyze available data needed to implement the sustainability initiatives Assess quality and identify gaps Collect missing data from internal and external sources	Harmonize the definitions and map internal with external reference data. Prepare and aggregate the data for further manipulations and calculations
Outcomes	Relevant data objects and attributes for the sustainability initiative, business analyst (sustainability report owner)	Quality assessment and gaps in existing data; collection of missing data objects and attributes from internal and external sources	Curated database for KPIs and sustainability reporting
Roles and responsibilities	Sustainability officer, compliance officer	Data steward, data analyst, business operations	Data steward, data engineer
Challenges	<ul style="list-style-type: none"> Difficulties in adapting to an increasing number of regulations and certifications that address the same SDGs Interpreting and translating the sustainability goals, legal texts, or certification label requirements into concrete data objects and attributes 	<ul style="list-style-type: none"> Inability to capture the necessary data along a global supply chain Missing or erroneous data (e.g., material description) which is presumed to be complete in the enterprise systems 	<ul style="list-style-type: none"> Heterogeneity of data sources (e.g., variability of types and formats between data from internal and external sources) Lack of definitions and semantics, as well as difficulties encountered when mapping against them (e.g., recycled material)

Table 42. Data sourcing practices for sustainability

Sense-making: This practice involves the time-consuming analysis of the sustainability goals, ambitions, and regulations and their interpretation in terms of data requirements. In initiatives where regulatory pressure is exerted, cross-functional teams – with sustainability, legal, and data expertise – must interpret the legal texts or lengthy certification contracts and translate them into tangible data requirements. In these instances, sense-making clarifies the data objects and attributes mentioned in the regulation – including a rough understanding of their semantics – which should be collected in the next data sourcing phase. In initiatives that were not a response to regulative pressures, but initiated by internal data management efforts, sense-making relates to translating ambitious sustainability goals and indicating how to measure them. The sense-making activities mainly focus on defining suitable measurement rules and understanding the data at hand, that is, identifying relevant data objects and attributes that are

already in the systems, as well as discovering gaps that must be filled to address the goals set by the sustainability initiatives. This implies that without exerted regulative pressures, enterprises pursue self-set goals and ambitions derived from insights based on available data (e.g., understanding the lower levels of granularity of product or packaging composition), being driven by the other two types of pressure.

Data collection: This practice involves the analysis of available data needed to realize the sustainability initiatives, quality assessment, and gap identification, as well as the collection of missing data from internal and external sources. We notice that the four sustainability initiatives mainly build on existing, well-defined data objects as input, which are repurposed for sustainability needs. For instance, *product master data* for the finished products, *material master data* for all parts, components, and raw materials, and the *BoM* are essential to gain an understanding of the composition of a finished product or its packaging at the lowest level of granularity. By contrast, there are new data objects which previously have not been maintained in companies' systems, and which must be created. These objects, among others, include specific KPIs (e.g., *plastic indicator*), as well as relevant meta information (e.g., *product label*, *certification body*, *regulation*). Even though several data objects already exist in companies' ERP and BI systems, sustainability activities repurpose and extend the scope of the data use (e.g., with amendments to the material descriptions and classifications) and, in turn, require the establishment of new business rules and data pipelines. Furthermore, sustainability initiatives often rely on external data that can only be collected from business partners or in terms of applicable industry benchmarks for environmental sustainability (e.g., SDG Ambition (SAP, 2020)). Based on the insights gained from the cases, we noted that the four initiatives rely on similar data objects and attributes, and we validated this learning in focus group sessions. We therefore decided to consolidate the data requirements in the form of a conceptual data model that supports sense-making, data collection, and data reconciliation practices. This model conceptualizes the data requirements with reference to ten relevant data objects and attributes of the identified sustainability initiatives (see Figure 20). A mutual understanding of these attributes prepares the ground for a common view of the initiatives' data requirements (see Table 44 in Appendix 2).

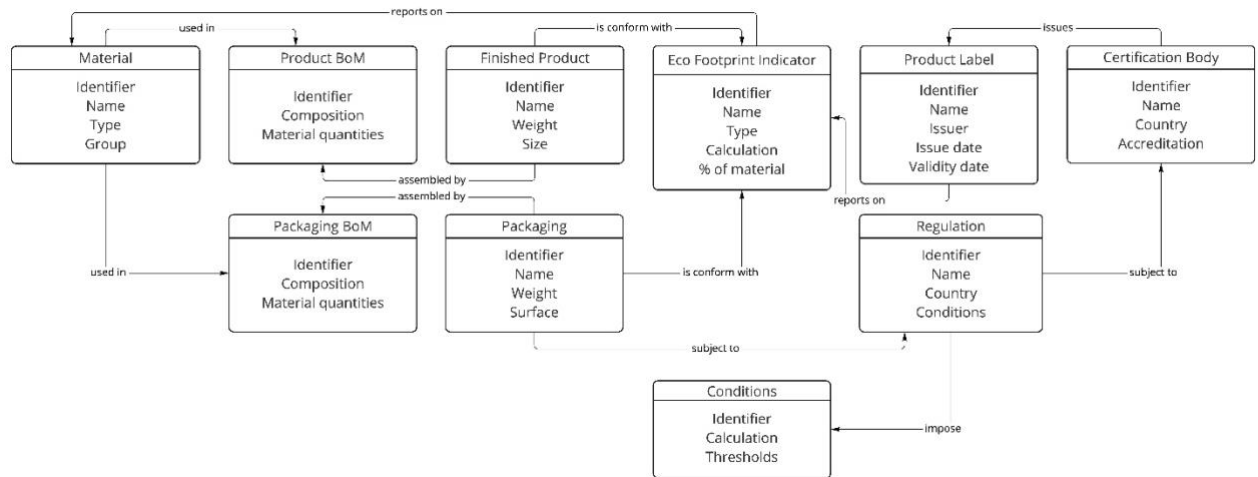


Figure 20: Data requirements for sustainability initiatives at product and packaging level

Data reconciliation: This third data sourcing practice encapsulates activities that prepare the data for sustainability reporting and harmonize data from different sources. For instance, to calculate the KPIs on the use of recycled material in a given product, internally and externally collected heterogeneous data should be brought together. Companies struggle with the variability of the sourced data’s types and formats, which complicates aggregation. For example, A faced difficulties in aggregating the data on components into the final product’s composition due to the different units of measurement (e.g., the surface and weight of the product). Finally, the required harmonization across heterogeneous data sources and the lack of definitions and semantics that cause difficulties in their use are among the most demanding challenges faced by companies.

6 Discussion, implications, and future action

6.1 Summary of contributions and discussion

Our study is an example of impact-oriented Green IS research (Gholami et al., 2016) that guides enterprises on their way to become more sustainable, while embedding sustainability in IS and in practice (Seidel et al., 2017). More specifically, our findings advance Green IS and EMIS literature that has, in the past, mainly identified issues concerning data quality and accessibility (Machado Ribeiro et al., 2022; Melville et al., 2017; Zampou et al., 2022) without elaborating on the data-related requirements and processes. To address this gap, our study introduces a data perspective on sustainability and draws attention to data sourcing practices as basis for reliable and trustworthy sustainability reporting. It makes a two-fold contribution: First, as theoretical contribution, we propose institutional theory as suitable lens to uncover how data sourcing practices are shaped in response to exerted external pressures. The resulting research framework allows to identify causal chains, leading from the relevant pressures to prioritized sustainability initiatives, and the emerging data sourcing practices. Second, our empirical contributions include novel, revelatory insights into key initiatives in the field of environmental sustainability, that touch on first, understanding the ecological footprint, and second, obtaining labels or complying with regulations, both on product and packaging levels. From our cross-case analysis, we derive three general data sourcing practices to address the data-related issues in sustainability initiatives: sense-making, data collection, and reconciliation. To support these three practices, we outline a conceptual data model that synthesizes the relevant data objects and attributes that need to be sourced for product-and packaging-related sustainability initiatives.

While we identify three general data sourcing practices, our findings also help to understand – via the lens of institutional theory – how the exerted pressures shape those practices: Interestingly, the normative and cognitive-cultural pressures were as prominent, if not more so, than the regulative pressures, in shaping the data sourcing practices. While firms of course gain and maintain legitimacy through attempting to navigate the complex regulations that have already emerged (Scott, 2013), we saw that the pressure from other organizations (Aldrich, 1979; DiMaggio & Powell, 1983) and customers played a substantial role in prioritizing sustainability initiatives and their data requirements. The context of sustainability transcends its regulative implications, where companies must vie for “political power, institutional legitimacy... as well as economic fitness” (DiMaggio and Powell 1983, p. 150) from its peers, competitors, and

customer constituents alike. Here we see that seeking legitimacy through conformity is not a static property, achieved only by creating data sourcing practices to simply comply with regulations and rules, but it seems to be instead a more dynamic process socially constructed by the companies (and regulators) (Burdon & Sorour, 2020; Suddaby et al., 2017). This implies that companies will continue to adjust and move forward with the three data sourcing practices that have thus far emerged as reactions to the emerging institutional pressures.

The data sourcing practices suggested in this study provide a basis for reliable and trustworthy sustainability reporting, thereby avoiding or mitigating the risks of greenwashing (Szabo & Webster, 2021). Interestingly, our findings also highlight that data sourcing for sustainability reporting is inherently more complex than for traditional reporting. In financial reporting, companies rely on established accounting standards and most of the data is generated internally and managed by accounting teams, whereas in the sustainability context, the requirements and responsibilities are yet to be clearly defined. Therefore, the sense-making derived from internal goals or regulations is an essential step to translate the high-level requirements from regulations or internal ambitions into concrete data requirements and identify data that should be sourced along the entire supply chain. Our study also reveals that data sourcing practices for sustainability rely on cross-functional collaboration between multiple stakeholders: sustainability and compliance officers as well as business analysts for sense-making; data stewards, data analysts, and business operations for data collection; and data stewards and data engineers for data reconciliation. The collaboration even goes beyond the internal stakeholders to embrace external parties, most importantly suppliers, logistics providers and other partners along the entire supply chain. Another characteristic of data sourcing for sustainability is that data must be repurposed (e.g., product or packaging dimensions), or even created on demand (e.g., prescribing the weight of recycled materials in a product). Thus, more heterogeneous data is collected from various (internal and external) sources, which must be integrated with internal systems and adapted for the new data and business requirements. This underpins that data reconciliation requires companies to develop integration and data management strategies that ensure seamless information flows and effective analytics.

6.2 Implications for research

Our research complements Green IS research, which has largely focused on EMIS design, in terms of components, features and design principles. It suggests adding a dedicated data perspective to this stream of research in order to address the data-related issues that have been highlighted in prior Green IS studies (Marx Gómez & Teuteberg, 2015; Melville et al., 2017;

Zampou et al., 2022). Our findings highlight that reporting on sustainability initiatives is not uniform across companies, but shaped in response to external pressures and goes hand in hand with the development of data sourcing practices. The identified practices come with challenges at different levels, which also represent interesting opportunities for future research – from the interpretation of sustainability-related regulations using formal or semi-formal approaches (sense-making), to the platforms supporting the gathering of data along global supply chains (data collection), and the definition of semantic data models in the form of knowledge graphs for sustainability-relevant information that allow to efficiently integrate data of heterogeneous formats and granularity (data reconciliation). While our study suggests ways to address the data-related issues that have been highlighted in prior Green IS studies (Marx Gómez & Teuteberg, 2015; Melville et al., 2017; Zampou et al., 2022), it also reveals that sustainability reporting becomes increasingly integrated into traditional corporate reporting. Thus, Green IS and EMIS researchers should study the disclosure requirements imposed by existing and emerging reporting regulations, such as CSRD, and investigate EMIS in the context of corporate reporting processes and platforms. The suggested research model can serve as a framework to theorize about CSRD as well as other industry- or country-specific sustainability regulations, their impact on sustainability initiatives and the development of data sourcing practices. It allows to identify patterns among different business contexts and settings and analyze the context-specificity of reporting requirements and data sourcing practices.

Our research also contributes to and has implications for the emerging body of research on data sourcing which extends prevailing IS/IT sourcing concepts by considering data as a specific object of sourcing (Jarvenpaa & Markus, 2020; Krasikov et al., 2022). Given the relevance of data sourcing in the context of sustainability, we call on the IS community to utilize this opportunity to further explore data sourcing practices “to reach an eventual symbiosis in which research informs practice and practice informs research” (Seidel et al., 2017, p. 42). Future research could use these findings to develop a holistic data sourcing theory in the context of enterprise-wide sustainability activities. We also see opportunities for academic research that explores how established data management principles and concepts complement data sourcing practices. Finally, the intersection of data sourcing and sustainability undeniably provides exciting opportunities for further inquiries into sustainable supply chains, Green IS, and sustainable computing, and the continued examination of EMIS purporting external data integration.

6.3 Implications for practice

Systematic data sourcing practices enable enterprises to accurately and transparently report on the progress of their sustainability initiatives. They do not only support compliance with the existing and upcoming reporting regulations, such as the European CSRD, but also help building trust with key stakeholders, most importantly their customers, and enhance the enterprise's reputation. For practitioners, our findings support companies that intend to go beyond ad-hoc approaches when fulfilling sustainability requirements and develop systematic data sourcing practices as a basis for reliable and trustworthy sustainability reporting. As the identified sustainability initiatives are of high relevance for many companies, practitioners can use the conceptual data model to identify typical data requirements for environmental sustainability and leverage the data sourcing practices as basis for setting up their internal data sourcing processes. Firstly, sense-making is an essential first step to translate regulations or internal goals and ambitions into concrete data requirements. Secondly, data collection focuses on identifying and gathering data (also data that has never been collected before), within and beyond the organizational boundaries, requiring external data from suppliers and other business partners. Thirdly, reconciliation of heterogenous sources is a challenging integration endeavor, which needs to be supported by ontologies and standards.

Our study highlights that even though enterprises are active in diverse industries and business contexts, reporting on environmental sustainability still requires them to report on the same data objects which are maintained in their ERP systems. Practitioners can use the conceptual data model to map data objects and attributes in these systems, to assess the need for enrichment from internal and external sources, and to define the target data model to reconcile data collected from heterogenous sources.

6.4 Limitations and outlook

Like most research, this study is not without limitations. First, it builds on empirical insights gleaned from the selected cases drawn from a larger pool of companies. While the identified challenges and practices are relevant for product- and packaging-related initiatives, they may not be generalizable to other contexts. Although we discussed the data sourcing challenges and practices in focus groups involving a larger group of companies that also prioritize other initiatives, our findings are limited to the scope of environmental sustainability initiatives. It would be interesting to replicate our study with initiatives in the fields of social and economic sustainability, thereby enlarging its generalization potential. Second, given that many companies are still in the early phases of their sustainability initiatives and that multiple

regulations are expected to be rolled out in future, there are opportunities for longitudinal studies that analyze the evolution of institutional pressures and data sourcing practices. While institutional theory offers valuable insights into the influence of external pressures on organizations' behavior and decision-making processes, it also has limitations. For instance, while it recognizes the importance of legitimacy, institutional theory may not fully account for ethical considerations related to data sourcing practices. Organizations may face conflicting pressures between achieving legitimacy and adhering to ethical principles, particularly in the context of sustainability. The theory's emphasis on conformity and legitimacy-seeking behavior may overshadow the ethical dimensions of data sourcing decisions. Furthermore, the theory often assumes a certain level of homogeneity in how organizations respond to external pressures, assuming conformity and isomorphism. This opens an interesting avenue for future research, namely observing the variation and diversity among organizations in their data sourcing strategies.

7 References

- Aldrich, H. (1979). *Organizations and Environments*. Prentice Hall, Englewood Cliffs, NJ.
- Bansal, P., & Roth, K. (2000). Why Companies Go Green: A Model of Ecological Responsiveness. *The Academy of Management Journal*, 43(4), 717–736.
- Benbasat, I., Goldstein, D. K., & Mead, M. (1987). The Case Research Strategy in Studies of Information Systems. *MIS Quarterly*, 11(3), 369–386.
- Bissinger, K., Brandi, C., Cabrera de Leicht, S., Fiorini, M., Schleifer, P., Fernandez de Cordova, S., & Ahmed, N. (2020). Linking Voluntary Standards to Sustainable Development Goals. International Trade Centre.
- Burdon, W. M., & Sorour, M. K. (2020). Institutional Theory and Evolution of ‘A Legitimate’ Compliance Culture: The Case of the UK Financial Service Sector. *Journal of Business Ethics*, 162(1), 47–80.
- Butler, T. (2011). Compliance with Institutional Imperatives on Environmental Sustainability: Building Theory on the Role of Green IS. *The Journal of Strategic Information Systems*, 20(1), 6–26.
- Castillo-Montoya, M. (2016). Preparing for Interview Research: The Interview Protocol Refinement Framework. *The Qualitative Report*, 21(5), 811–831.
- Christensen, H. B., Hail, L., & Leuz, C. (2021). Mandatory CSR and Sustainability Reporting: Economic Analysis and Literature Review. *Review of Accounting Studies*, 26(3), 1176–1248.
- Competition and Markets Authority. (2021). Global Sweep finds 40% of Firms’ Green Claims Could be Misleading. GOV.UK. <https://www.gov.uk/government/news/global-sweep-finds-40-of-firms-green-claims-could-be-misleading>
- Corbett, J., & Mellouli, S. (2017). Winning the SDG Battle in Cities: How an Integrated Information Ecosystem Can Contribute to the Achievement of the 2030 Sustainable Development Goals. *Information Systems Journal*, 27(4), 427–461.
- Daddi, T., Bleischwitz, R., Todaro, N. M., Gusmerotti, N. M., & De Giacomo, M. R. (2020). The Influence of Institutional Pressures on Climate Mitigation and Adaptation Strategies. *Journal of Cleaner Production*, 244, 118879.
- Deloitte. (2021). The Importance of ESG Data Management: Challenges and Opportunities for the Real Estate Ecosystem. <https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/financial-services/deloitte-nl-fsi-real-estate-esg-data-management-whitepaper.pdf>
- DiMaggio, P. J., & Powell, W. W. (1983). The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields. *American Sociological Review*, 48(2), 147–160.
- EDM Council. (2022). ESG Data Management: Asset Owners. <https://edmcouncil.org/groups-leadership-forums/esg-data-management/>
- El-Gayar, O., & Fritz, B. D. (2006). Environmental Management Information Systems (EMIS) for Sustainable Development: A Conceptual Overview. *Communications of the Association for Information Systems*, 17(1), 756–784.
- European Commission. (2021). Screening of Websites for ‘Greenwashing’: Half of Green Claims Lack Evidence. https://ec.europa.eu/commission/presscorner/detail/en/ip_21_269
- European Commission. (2022). A European Green Deal. European Commission - European Commission. https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_en
- Frost, G., Jones, S., & Lee, P. (2012). *The Measurement and Reporting of Sustainability Information within the Organization: A Case Analysis*. Emerald Group Publishing.
- Galleli, B., Semperebon, E., Santos, J. A. R. dos, Teles, N. E. B., Freitas-Martins, M. S. de, & Onevetch, R. T. da S. (2021). Institutional Pressures, Sustainable Development Goals and COVID-19: How Are Organisations Engaging? *Sustainability*, 13(21), 12330.
- Gholami, R., Watson, R., Molla, A., Hasan, H., & Bjorn-Andersen, N. (2016). Information Systems Solutions for Environmental Sustainability: How Can We Do More? *Journal of the Association for Information Systems*, 17(8), 521–536.
- Glover, J. L., Champion, D., Daniels, K. J., & Dainty, A. J. (2014). An Institutional Theory Perspective on Sustainable Practices Across the Dairy Supply Chain. *International Journal of Production Economics*, 152, 102–111.
- GRI. (2022). A Short Introduction to the GRI standards. <https://www.globalreporting.org/media/wtafi4tw/a-short-introduction-to-the-gri-standards.pdf>
- Hilpert, H., Kranz, J., & Schumann, M. (2014). An Information System Design Theory for Green Information Systems for Sustainability Reporting—Integrating Theory with Evidence from Multiple Case studies. *Proceedings of the 22nd European Conference on Information Systems*.
- Ivan, O. R. (2009). Sustainability in Accounting-basis: A Conceptual Framework. *Annales Universitatis Apulensis: Series Oeconomica*, 11(1), 106–116.
- Jarvenpaa, S. L., & Markus, M. L. (2020). Data Sourcing and Data Partnerships: Opportunities for IS Sourcing Research. In *Information Systems Outsourcing* (5th ed., pp. 61–79). Springer.
- Krasikov, P., Eurich, M., & Legner, C. (2022). Unleashing the Potential of External Data: A DSR-based Approach to Data Sourcing. *Proceedings of the 30th European Conference on Information Systems*.

- Lu, Y., Zhao, C., Xu, L., & Shen, L. (2018). Dual Institutional Pressures, Sustainable Supply Chain Practice and Performance Outcome. *Sustainability*, 10(9), 3247.
- Machado Ribeiro, V., Barata, J., & Cunha, P. da. (2022). Sustainable Data Governance: A Systematic Review and a Conceptual Framework. *Proceedings of the 30th International Conference on Information Systems Development*.
- Marx Gómez, J., & Teuteberg, F. (2015). Toward the Next Generation of Corporate Environmental Management Information Systems: What is Still Missing? In L. M. Hilty & B. Aebischer (Eds.), *ICT Innovations for Sustainability* (pp. 313–332). Springer International Publishing.
- Melville, N. (2010). Information Systems Innovation for Environmental Sustainability. *Management Information Systems Quarterly*, 34(1), 1–21.
- Melville, N., Saldanha, T. J. V., & Rush, D. E. (2017). Systems Enabling Low-Carbon Operations: The Salience of Accuracy. *Journal of Cleaner Production*, 166, 1074–1083.
- Miles, M. B., Michael, H. A., & Johnny, S. (2014). *Qualitative Data Analysis: A Methods Sourcebook* (3rd ed.). SAGE Publications.
- Milne, M. J., & Gray, R. (2013). W (h)ither Ecology? The Triple Bottom Line, the Global Reporting Initiative, and Corporate Sustainability Reporting. *Journal of Business Ethics*, 118, 13–29.
- Nuthi, K. (2022). An EU Open Data Plan Can Help Combat Climate Change. Center for Data Innovation. <https://datainnovation.org/2022/08/an-eu-open-data-plan-can-help-combat-climate-change/>
- Official Journal of the European Union. (2014). Directive 2014/95/EU. <http://data.europa.eu/eli/dir/2014/95/oj/eng>
- Official Journal of the European Union. (2022, December 14). Directive (EU) 2022/2464. <http://data.europa.eu/eli/dir/2022/2464/oj/eng>
- Pan, S. L., Carter, L., Tim, Y., & Sandeep, M. S. (2022). Digital Sustainability, Climate Change, and Information Systems Solutions: Opportunities for Future Research. *International Journal of Information Management*, 63, 102444.
- Powell, W. W., & DiMaggio, P. J. (2012). *The New Institutionalism in Organizational Analysis*. University of Chicago press.
- Raj, A., Agrahari, A., & Srivastava, S. K. (2020). Do Pressures Foster Sustainable Public Procurement? An Empirical Investigation Comparing Developed and Developing Economies. *Journal of Cleaner Production*, 266, 122055.
- SAP. (2020). UN, Accenture and SAP Launch SDG Ambition Guides. <https://news.sap.com/2020/09/sdg-ambition-guides-un-global-compact-accenture-sap-3m/>
- Scott, W. R. (2013). *Institutions and Organizations: Ideas, Interests, and Identities* (4th ed.). SAGE Publications.
- Seethamraju, R., & Frost, G. (2016). Information Systems for Sustainability Reporting—A State of Practice. *Proceedings of the 22nd Americas Conference on Information Systems*.
- Seidel, S., Bharati, P., Fridgen, G., Watson, R., Albizri, A., Boudreau, M.-C., Butler, T., Kruse, L., Guzman, I., Karsten, H., Lee, H., Melville, N., Rush, D., Toland, J., & Watts, S. (2017). The Sustainability Imperative in Information Systems Research. *Communications of the Association for Information Systems*, 40(1), 40–52.
- Sisaye, S. (2021). The Organizational Ecological Resource Framework of Sustainability Reporting: Implications for Corporate Social Reporting (CSR). *Journal of Business and Socio-Economic Development*, 2(2), 99–116.
- Stindt, D., Nuss, C., Bensch, S., Dirr, M., & Tuma, A. (2014). An Environmental Management Information System for Closing Knowledge Gaps in Corporate Sustainable Decision-Making. *Proceedings of the 35th International Conference on Information Systems*.
- Stoll, A. (2022). ESG in Your Value Chain. KPMG. <https://home.kpmg/ch/en/home/insights/2022/06/esg-supply-chain.html>
- Suddaby, R., Bitektine, A., & Haack, P. (2017). Legitimacy. *Academy of Management Annals*, 11(1), 451–478.
- Szabo, S., & Webster, J. (2021). Perceived Greenwashing: The Effects of Green Marketing on Environmental and Product Perceptions. *Journal of Business Ethics*, 171(4), 719–739.
- Teo, H. H., Wei, K. K., & Benbasat, I. (2003). Predicting Intention to Adopt Interorganizational Linkages: An Institutional Perspective. *MIS Quarterly*, 27(1), 19–49.
- Teuteberg, F., & Straßenburg, J. (2009). State of the Art and Future Research in Environmental Management Information Systems—A Systematic Literature Review. In *Information Technologies in Environmental Engineering* (pp. 64–77).
- Tremblay, D., Gowsy, S., Riffon, O., Boucher, J.-F., Dubé, S., & Villeneuve, C. (2021). A Systemic Approach for Sustainability Implementation Planning at the Local Level by SDG Target Prioritization: The Case of Quebec City. *Sustainability*, 13(5), 1–20.
- United Nations. (2022). The 17 Goals | Sustainable Development. <https://sdgs.un.org/goals>
- Van de Ven, A. H., & Poole, M. S. (2005). Alternative Approaches for Studying Organizational Change. *Organization Studies*, 26(9), 1377–1404.
- Vrolijk, H. C. J., Poppe, K. J., & Keszthelyi, S. (2016). Collecting Sustainability Data in Different Organisational Settings of the European Farm Accountancy Data Network. *Studies in Agricultural Economics*, 118(3), 138–144.
- Walls, J. L., Phan, P. H., & Berrone, P. (2011). Measuring Environmental Strategy: Construct Development, Reliability, and Validity. *Business & Society*, 50(1), 71–115.
- Wang, X., Brooks, S., & Sarker, S. (2015). A Review of Green IS Research and Directions for Future Studies. *Communications of the Association for Information Systems*, 37(1), 395–429.

- Watson, R. T., Boudreau, M.-C., & Chen, A. J. (2010). Information Systems and Environmentally Sustainable Development: Energy Informatics and New Directions for the IS Community. *MIS Quarterly*, 34(1), 23–38.
- Yang, C.-S. (2018). An Analysis of Institutional Pressures, Green Supply Chain Management, and Green Performance in the Container Shipping Context. *Transportation Research Part D: Transport and Environment*, 61, 246–260.
- Yin, R. K. (2009). *Case Study Research: Design and Methods* (4th ed.). SAGE Publications.
- Zampou, E., Mourtos, I., Pramataris, K., & Seidel, S. (2022). A Design Theory for Energy and Carbon Management Systems in the Supply Chain. *Journal of the Association for Information Systems*, 23(1), 329–372.

Appendix 1

Company	Job title	Years of experience	Interview duration	Industry	Company's revenue / employees	Key sustainability initiatives
A	Director data governance	19 years (7 years in company A)	78 minutes	Fashion and retail	\$1B-50B / ~60,000	Product labeling
B	Director master data management	15 years (11 years in company B)	75 minutes	Engineering and electronics	\$1B-50B / ~400,000	Product ecological footprint
C	Head of product data management	20 years (20 years in company C)	63 minutes	Pharmaceutical, chemicals	\$1B-\$50B / ~100 000	Product labeling
D	Data steward material & product	10 years (3 years in company D)	59 minutes	Manufacturing, chemicals	\$1B-\$50B / ~5,000	Plastic packaging tax
E	Global master data lead	27 years (16 years in company E)	58 minutes	Consumer goods	\$50B-\$100B / ~350,000	Packaging recyclability

Table 43. Interviewee profiles

Appendix 2

Data object	Attribute	Definition
Material	Identifier	A unique identifier assigned to the material
	Name	States the name assigned to the material and defines the material
	Type	Specifies material categorized together and defines the available views on the material
	Group	Classifies a group of materials with similar attributes and specifies the use of this group
Product BoM	Identifier	A unique identifier assigned to the product's bill of material
	Composition	Specifies the materials used to manufacture the product
	Material quantities	Specifies the quantities of used materials
Packaging BoM	Identifier	A unique identifier assigned to the packaging's bill of material
	Composition	Specifies the materials used in the manufactured packaging
	Material quantities	Specifies the quantities of used materials
Finished Product	Identifier	A unique identifier assigned to the product
	Name	States the name assigned to the finished product and defines the product
	Weight	Specifies the finished product's weight
	Size	Specifies the finished product's size
Packaging	Identifier	A unique identifier assigned to the packaging
	Name	States the name assigned to the packaging and defines the packaging
	Weight	Specifies the packaging's weight
	Surface	Specifies the packaging's surface
Eco-footprint indicator	Identifier	A unique identifier assigned to the ecological footprint indicator
	Name	States the name assigned to the indicator and defines the indicator
	Type	Specifies indicators categorized together and defines the available views on the indicator
	Calculation	Defines the calculation rules for the indicator
	% of material	Specifies the quantities of materials used in the calculation
Product Label	Identifier	A unique identifier assigned to the product label
	Name	States the name assigned to the product label and defines the product label
	Issuer	Name of the issuing organization for the product label
	Issue date	Date on which the product label was issued
	Validity date	Date until which the product label is valid
Certification Body	Identifier	A unique identifier assigned to the certification body
	Name	States the name of the certification body
	Country	The country in which the certification body is located
	Accreditation	Confirms the competence of the certification body according to internationally recognized standards
Regulation	Identifier	A unique identifier assigned to the regulation
	Name	States the name assigned to the regulation and defines the regulation
	Country	The country (or countries) in which the regulation is applicable
	Condition	Conditions imposed by the underlying regulations
Condition	Identifier	A unique identifier assigned to the condition
	Calculation	Defines the calculation rules for the required compliance regulation
	Thresholds	Defined thresholds in accordance with the regulatory requirements

Table 44. List of definitions of the attributes