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

A Method to Screen, Assess, and Prepare Open Data

PAVEL KRASIKOV

Faculty of Business and Economics (HEC), University of Lausanne, Switzerland, pavel.krasikov@unil.ch

CHRISTINE LEGNER

Faculty of Business and Economics (HEC), University of Lausanne, Switzerland, christine.legner@unil.ch

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.

CCS CONCEPTS • Information systems ~ World Wide Web ~ Web searching and information discovery

Additional Keywords and Phrases: Open Data, Data Preparation, Data Quality, Action Design Research, Knowledge Graph

1 INTRODUCTION

Open data is known to be free for use, reuse, and redistribution by anyone [46]. It offers business and innovation potential to companies and national economies [35,70], with an estimated total market size in the European Union of 325 billion euros [27]. 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 [70,57,24]. 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 [25] and lead to a "mismatch between the needs and expectations of the users and the possibilities offered by available datasets" [55], 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 [7,37,61]. To address these issues, researchers have developed dedicated assessment techniques, such as the "Luzzu" framework [23], the "LANG" approach [66], or the "QUIN" usability criteria [48]. 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 [67]. To

1936-1955/2023/1-ART1 $15.00
http://dx.doi.org/10.1145/3603708

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 [58], 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 [36] 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 [47], with data.europa.eu and data.gov combined providing access to more than 1.7 million open datasets [21,26]. Despite these impressive numbers, open data use by enterprises remains below expectations

[70]. 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” [15]. One of the major misconceptions about open data [35] 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 [68,13,5,65,66], but they remain insufficient and have been criticized in terms of functionalities, namely in the public sector [19,40]. Although open data literature has identified a large set of barriers [35], three main categories stand out as barriers for the enterprise use of open data [37]: 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” [35], 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 [41,68]. 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 [35,37]. Table 1 synthetizes the main categories of barriers and their impact on enterprises as open data consumers.

Table 1: Main barriers to open data adoption in enterprises and their impact on open data consumption

Category	Description	Impact on enterprises	Sources
Transparency	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 [35]; Zuiderwijk et al., 2012 [68]

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 [35];
			Zuiderwijk et al., 2012 [68];
			Martin et al., 2013 [41];
			Conradie and Choenni, 2014 [18];
			Barry and Bannister, 2014 [8]

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 [35];
			Zuiderwijk et al., 2012 [68];
			Conradie and Choenni, 2014 [18];
			Beno et al., 2017 [11];
			Corsar and Edwards, 2017 [19]

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 2), 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 [23,61,66] are inspired by classical methodologies on data quality assessment, especially those proposed by Batini et al. [9] and Pipino et al. [51]. 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 [19]. Hence, although the open data assessment literature provides a clear link to the traditional data quality literature [66], it neglects the open data consumers' perspective [38]. We argue that the definition of data quality, commonly referred to as "fitness for use" [62], must equally apply to open data, emphasizing the importance of open data's "usefulness" in specific use cases [48], 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.

Table 2: Open data assessment techniques

Source	Assessment approach	Assessment scope	Methodology
Bogdanović-Dinić et al., 2014 [14]	"Data openness score" based on eight open data principles [45]	Metadata	Case study: application of the "data openness" model to 7 open data portals
Reiche et al., 2014 [53]	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 [23]	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 [44]	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 [61]	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 [39]	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 [63]	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 [48]	Usability evaluation "QUIN" criteria (12 usability criteria)	Platform features	Evaluation as part of the agile development process "ROUTE-TO-PA"
Bicevskis et al., 2018 [12]	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
Stróżyna et al., 2018 [60]	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 [66]	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 [43]	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

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 3). Only two of the existing studies address the processes exclusively from the consumers' perspective [30,70]. 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 [54] 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 [69] 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" [69].

Table 3: Publishers' and consumers' perspectives on open data processes (based on [36])

Source	Perspective and context	Research method	Processes (publishers)	Processes (consumers)
Ren and Glissmann, 2012 [54]	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
Masip-Bruin et al, 2013 [42]	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 [69]	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 [30]	Big data user Integration techniques for structured and unstructured online (open) data	Explorative analysis	N/A	Discover, integrate, and validate open datasets
Zuiderwijk et al., 2015 [70]	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 [20]	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 [1]	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 [2]	Open data publisher Integrating and publishing linked open government data	Illustrative case study	Data transformation, interlinking, storage, visualization, and publishing	N/A

Continuing the exploration of open data barriers, Crusoe and Melin [20] expand the open government data process [69], 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 [16,33,34] 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. [70] 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 [30] distinguishes between three major steps in the use of heterogeneous online datasets: discovery, integration, and validation. Finally, Abella et al. [1] 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 [25]. 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 [48], 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 [50]. 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 [58]. In contrast to existing DS methods that relegate evaluation to a subsequent phase, ADR incorporates evaluation into the design cycles [58]. It allows to create rigorous and relevant business knowledge that will help to develop "specific solution(s) in specific situation(s)" [4] and learn from the instantiations. The outcome of our research is categorized as a method that explains "what to do in different situations" [28] in accordance with a stepwise structure, while also including additional constituents such as notation, procedural guidelines, and concepts [56], 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 [29] 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 [32]. 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 [58]. 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 4 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 heterogenous 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.

Table 4: BIE cycles and their contribution to method development

	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: • Development of the method's phases 1 to 3 • Focus on Phase 3 (preparation for use)	Beta version of the method: • 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

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 [52]. 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" [63]. 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 [38].
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 [51,62]. 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 [70,13,5,65]. The conceptualization of the domain of interest through ontologies is a known solution when it comes to the integration of large and unknown datasets [17]. 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" [22]. 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" [49].

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 6 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 [10,57,59]. 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 5 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.

Table 5: Example of use case ideation

Use case category and example	Description	Open datasets and providers	Internal data objects	Data management impact	Business impact
Data management use case: 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
Business process use case: 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
Analytics and intelligence use case: 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 6: Overview of the method to screen, assess, and prepare open data for use

	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	• 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	• 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 source(s) that fit the purpose of the use case • Define relevant business concepts for the use case as a reference ontology	• 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	• Assess schema completeness for the required attributes predefined for the use case • Analyze the presence of the required attributes for use case feasibility	• Assess content quality based on the applicable data quality dimensions	• Provide full metadata documentation, including access, licensing, provenance, and publisher details • Document the dataset attributes	• 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

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 7 presents examples of identified datasets for corporate registers from leading EU countries in the open data initiatives [31] and leading world economies with recognized open data initiatives [71,72], 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 [73], 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.

Table 7: Example of identified open data sources and datasets

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

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 8), suggested by the comprehensive approaches of Neumaier et al. [44], Vetrò et al. [61], and Zhang et al. [66]. 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 [61,66], 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 [9]. 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).

Table 8: Relevant open data quality dimensions on metadata, schema and dataset content level (based on [44,61,66])

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).

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 [44] metadata quality assessment framework suggests the verification of the existence of metadata attributes of Data Catalog Vocabulary (DCAT) [3] 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 [51], 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" [38]. 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" [51]. 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 9), 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 [61,66], 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 8. 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 7) and the business partner data curation use case, as formulated in Phase 0. To perform these activities, we used the Pandas Profiling [64] 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 9). 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 9). 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.

Table 9: Examples of the datasets’ assessment results on the metadata, schema, and content levels

Dataset	Metadata	Schema	Content
Argentinian National Registry of Companies *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.*	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 *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.*	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 *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.*	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 *Observation: Although certain details are*	Identification: RA000423 Country: Latvia Format: CSV, XSLX Access login: no Free lookup service: available	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 Attribute (country): 0% missing

Dataset	Metadata	Schema	Content
missing in the metadata, the dataset is maintained with a comparably high level of quality.	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 (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 *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.*	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”	Total missing cells (all dataset): 37.2% Total duplicate rows (all dataset): 0.0% 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 *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.*	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”	Total missing cells (all dataset): 54.0% Total duplicate rows (all dataset): 0.0% 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 *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.*	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”	Total missing cells (all dataset): 50.9% Total duplicate rows (all dataset): 0.0% 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

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 noticable 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.

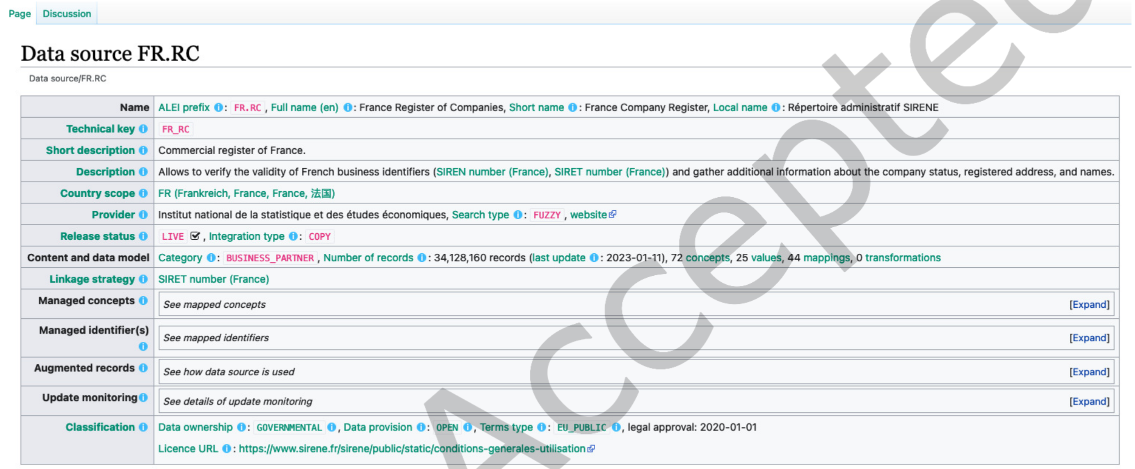
Phase 3 – preparation for use.

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 [6]. 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 [70,13,5,65]. 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 4), a productive documentation is maintained using a MediaWiki with an extension of Semantic MediaWiki. Figure 1 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 2).



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

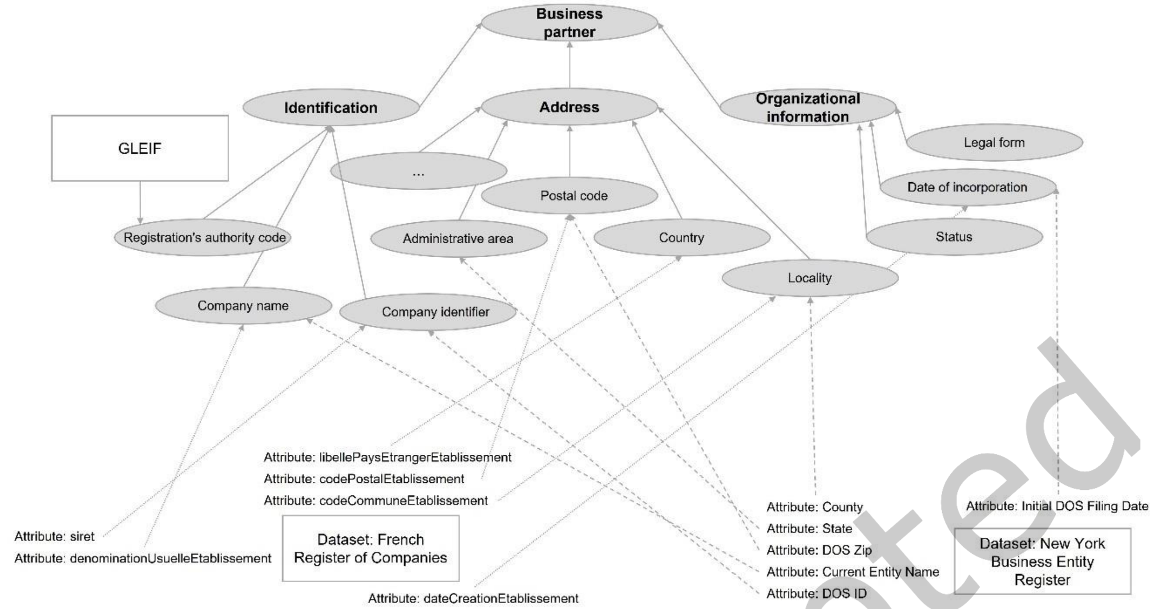

Business partner
Identification
Address
Organizational information
GLEIF
...
Postal code
Legal form
Date of incorporation
Registration's authority code
Administrative area
Country
Status
Locality
Company name
Company identifier
Attribute: libellePaysEtrangerEtablissement
Attribute: codePostalEtablissement
Attribute: codeCommuneEtablissement
Attribute: siret
Attribute: denominationUsuelleEtablissement
Dataset: French Register of Companies
Attribute: dateCreationEtablissement
Attribute: County
Attribute: State
Attribute: DOS Zip
Attribute: Current Entity Name
Attribute: DOS ID
Attribute: Initial DOS Filing Date
Dataset: New York Business Entity Register

Figure 2: 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 application of the method in practice can be non-linear. To illustrate, Figure 3 presents a workflow and thereby highlights the variations and sequencing of our method's applications in enterprises beyond the main flow illustrated in the previous sections, thus allowing for process flexibility and adaptability.

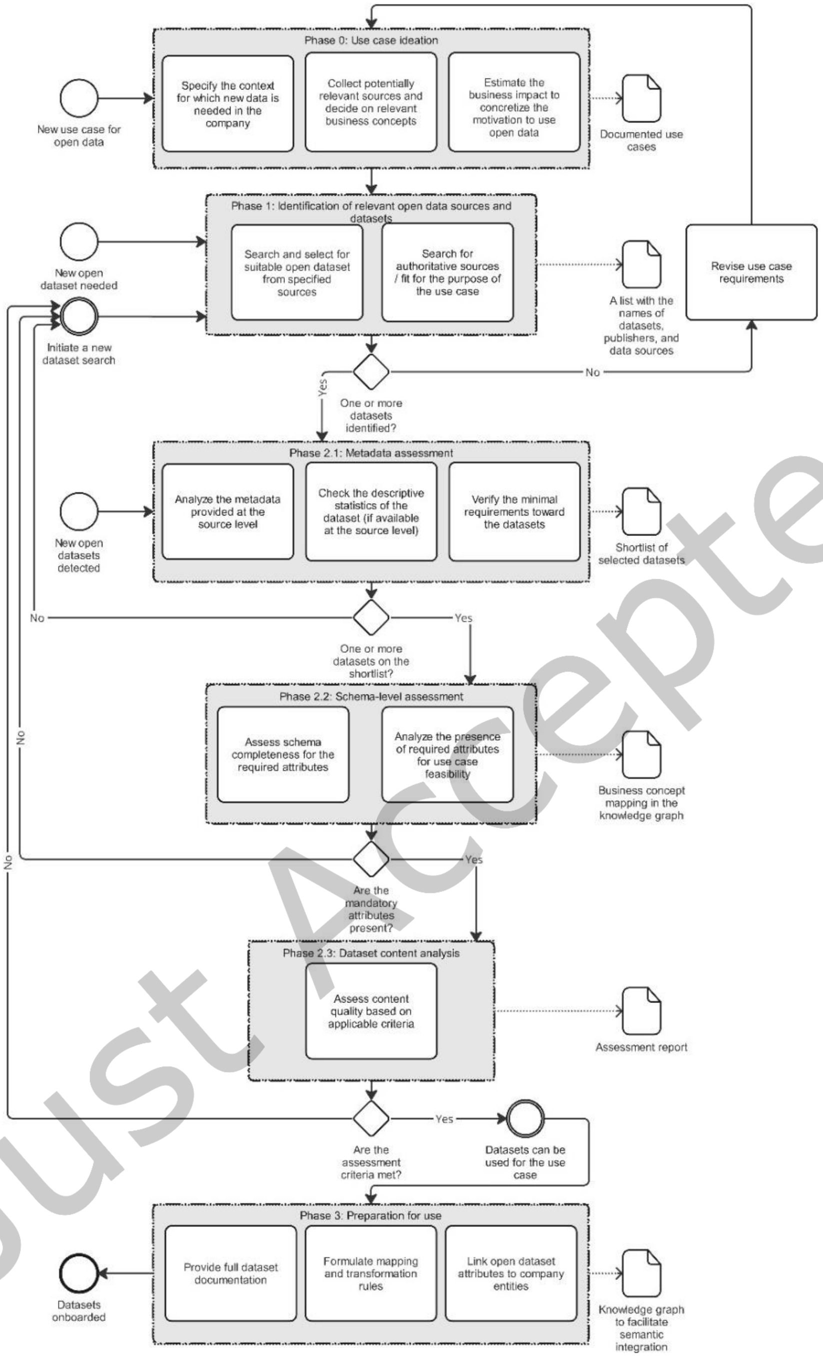

Phase 0: Use case ideation
New use case for open data
Specify the context for which new data is needed in the company
Collect potentially relevant sources and decide on relevant business concepts
Estimate the business impact to concretize the motivation to use open data
Documented use cases
Phase 1: Identification of relevant open data sources and datasets
New open dataset needed
Initiate a new dataset search
Search and select for suitable open dataset from specified sources
Search for authoritative sources / fit for the purpose of the use case
A list with the names of datasets, publishers, and data sources
Revise use case requirements
Yes
No
One or more datasets identified?
Phase 2.1: Metadata assessment
New open datasets detected
Analyze the metadata provided at the source level
Check the descriptive statistics of the dataset (if available at the source level)
Verify the minimal requirements toward the datasets
Shortlist of selected datasets
No
Yes
One or more datasets on the shortlist?
Phase 2.2: Schema-level assessment
Assess schema completeness for the required attributes
Analyze the presence of required attributes for use case feasibility
Business concept mapping in the knowledge graph
No
No
Yes
Are the mandatory attributes present?
Phase 2.3: Dataset content analysis
Assess content quality based on applicable criteria
Assessment report
Yes
Are the assessment criteria met?
Datasets can be used for the use case
Phase 3: Preparation for use
Datasets onboarded
Provide full dataset documentation
Formulate mapping and transformation rules
Link open dataset attributes to company entities
Knowledge graph to facilitate semantic integration

Figure 3. Possible variations in the workflow of the method to screen, assess, and prepare open data for use

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 scope, 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 meta-data 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.

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 10 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.

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 Stróżyna et al.'s quality-based selection framework [60] 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 holistic open data assessment framework [63] 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 metadata quality assessment framework [44], Vetrò et al.'s measurement framework [61], and Zhang et al.'s "LANG" approach [66], that were used in Phase 2 (see subsection 4.2), as they 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.

Table 10. Comparison with other approaches

Approach	Purpose	Screen	Assess		Prepare for use (integration)
			Use context awareness	Scope	
"Luzzu" framework [23]	Linked data quality assessment	no	no	Metadata and dataset	yes
Metadata quality assessment framework [44]	Automated metadata quality assessment for various open data portals	no	no	Metadata	no
Measurement framework [61]	Quantitative assessment of open government data quality	no	no	Metadata and dataset	no
Benchmarking framework [39]	Evaluation of open data portals' quality	no	no	Metadata	no
Holistic open data assessment framework [63]	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 [60]	Selection of open data sources to be fused with internal data	yes	yes	Metadata	yes
"LANG" approach [66]	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**

6 CONCLUSION, DISCUSSION, 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" [63]. 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 [70,13,5,65].

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 [25]. 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 [25]. The suggested method also demonstrates how semantic technologies, resulting from technical open data research streams [22,43], 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.

REFERENCES

[1] Alberto Abella, Marta Ortiz-de-Urbina-Criado, and Carmen De-Pablos-Heredero. 2019. The Process of Open Data Publication and Reuse. J. Assoc. Inf. Sci. Technol. 70, 3 (February 2019), 296–300. https://doi.org/10.1002/asi.24116

[2] Rabeb Abida, Emna Hachicha Belghith, and Anthony Cleve. 2020. An End-to-End Framework for Integrating and Publishing Linked Open Government Data. In Proceedings of the 29th International Conference on Enabling Technologies (WETICE'20), IEEE, Bayonne, France, 257–262. https://doi.org/10.1109/WETICE49692.2020.00057

[3] Riccardo Albertoni, David Browning, Simon Cox, Alejandra Gonzalez Beltran, Andrea Perego, and Peter Winstanley. 2020. Data Catalog Vocabulary (DCAT). Retrieved April 22, 2022 from https://www.w3.org/TR/vocab-dcat/

[4] Daniel Andriessen. 2008. Combining Design-Based Research and Action Research to Test Management Solutions. In Towards Quality Improvement of Action Research: Developing Ethics and Standard. Brill, 125–134.

[5] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In The Semantic Web (Lecture Notes in Computer Science), Springer, Berlin, Heidelberg, 722–735. https://doi.org/10.1007/978-3-540-76298-0_52

[6] Sören Auer, Volha Bryl, and Sebastian Tramp (Eds.). 2014. *Linked Open Data - Creating Knowledge Out of Interlinked Data*. Springer, Cham.

[7] Arief Bachtiar, Suhardi, and Wardani Muhamad. 2020. Literature Review of Open Government Data. In Proceedings of the 2020 International Conference on Information Technology Systems and Innovation (ICITSI'20), IEEE, Bandung, Indonesia, 329–334. https://doi.org/10.1109/ICITSI50517.2020.9264960

[8] Emily Barry and Frank Bannister. 2014. Barriers to Open Data Release: A View from the Top. IP 19, 1/2 (June 2014), 129–152. https://doi.org/10.3233/IP-140327

[9] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. 2009. Methodologies for Data Quality Assessment and Improvement. ACM Comput. Surv. 41, 3 (July 2009), 1–52. https://doi.org/10.1145/1541880.1541883

[10] Nicolas Baud, Antoine Frachot, and Thierry Roncalli. 2002. Internal Data, External Data and Consortium Data - How to Mix Them for Measuring Operational Risk. SSRN (June 2002), 1–18. https://doi.org/10.2139/ssrn.1032529

[11] Martin Beno, Kathrin Figl, Jurgen Umbrich, and Axel Polleres. 2017. Open Data Hopes and Fears: Determining the Barriers of Open Data. In Proceedings of the 2017 Conference for E-Democracy and Open Government (CeDEM'17), IEEE, Krems, Austria, 69–81. https://doi.org/10.1109/CeDEM.2017.22

[12] Janis Bicevskis, Zane Bicevska, Anastasija Nikiforova, and Ivo Oditis. 2018. Data Quality Evaluation: A Comparative Analysis of Company Registers' Open Data in Four European Countries. In Communication Papers of the 2018 Federated Conference on Computer Science and Information Systems (FedCSIS), Poznań, Poland, 197–204. https://doi.org/10.15439/2018F92

[13] Christian Bizer, Tom Heath, and Tim Berners-Lee. 2009. Linked Data - The Story So Far. Int. J. Semantic Web Inf. Syst. 5, 3 (July 2009), 1–22. https://doi.org/10.4018/jswis.2009081901

[14] Sanja Bogdanović-Dinić, Nataša Veljković, and Leonid Stoimenov. 2014. How Open Are Public Government Data? An Assessment of Seven Open Data Portals. In Measuring E-government Efficiency. Springer, New York, 25–44. https://doi.org/10.1007/978-1-4614-9982-4_3

[15] Katrin Braunschweig, Julian Eberius, Maik Thiele, and Wolfgang Lehner. 2012. The State of Open Data: Limits of Current Open Data Platforms. In Proceedings of the 21st International Conference on World Wide Web (WWW'12), ACM, Lyon, France.

[16] Anamaria Buda, Jolien Ubacht, and Marijn Janssen. 2016. Decision Support Framework for Opening Business Data. In Proceedings of the 16th European Conference on e-Government (ECEG 2016), Kidmore End: Academic Conferences International Limited, Ljubljana, Slovenia, 29–37.

[17] Tiziana Catarci, Monica Scannapieco, Marco Console, and Camil Demetrescu. 2017. My (Fair) Big Data. In Proceedings of the 2017 IEEE International Conference on Big Data (Big Data'17), IEEE, Boston, MA, USA, 2974–2979. https://doi.org/10.1109/BigData.2017.8258267

[18] Peter Conradie and Sunil Choenni. 2014. On the Barriers for Local Government Releasing Open Data. Government Information Quarterly 31, 1 (June 2014), S10–S17. https://doi.org/10.1016/j.giq.2014.01.003

[19] David Corsar and Peter Edwards. 2017. Challenges of Open Data Quality: More Than Just License, Format, and Customer Support. J. Data and Information Quality 9, 1 (March 2017), 1–4. https://doi.org/10.1145/3110291

[20] Jonathan Crusoe and Ulf Melin. 2018. Investigating Open Government Data Barriers: A Literature Review and Conceptualization. In Electronic Government. Springer, Cham, 169–183. https://doi.org/10.1007/978-3-319-98690-6_15

[21] Data.gov. 2022. Data.gov. Retrieved April 21, 2022 from https://www.data.gov/

[22] Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, Antonella Poggi, and Riccardo Rosati. 2018. Using Ontologies for Semantic Data Integration. In A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years. Springer, Cham, 187–202. https://doi.org/10.1007/978-3-319-61893-7_11

[23] Jeremy Debattista, Sören Auer, and Christoph Lange. 2016. Luzzu - A Methodology and Framework for Linked Data Quality Assessment. J. Data and Information Quality 8, 1 (November 2016), 1–32. https://doi.org/10.1145/2992786

[24] Tobias Enders, Carina Benz, and Gerhard Satzger. 2021. Untangling the Open Data Value Paradox. In Proceedings of the 16th International Conference on Wirtschaftsinformatik (WI'21), Springer, Cham, 200–205. https://doi.org/10.1007/978-3-030-86800-0_15

[25] Tobias Enders, Carina Benz, Ronny Schüritz, and Pamela Lujan. 2020. How to Implement an Open Data Strategy? Analyzing Organizational Change Processes to Enable Value Creation by Revealing Data. In Proceedings of the 28th European Conference on Information Systems (ECIS'20), An Online AIS Conference.

[26] EU Open Data Portal. 2022. EU Open Data Portal. Retrieved April 21, 2022 from https://data.europa.eu/en

[27] European Commission, Capgemini Consulting, Intrasoft International, Fraunhofer Fokus, con.terra, Sogeti, Open Data Institute, Time.lex, and 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, Luxembourg, Belgium.

[28] Göran Goldkuhl, Mikael Lind, and Ulf Seigerroth. 1998. Method Integration: The Need for a Learning Perspective. In IEE Proceedings - Software, Springer, 113–118. https://doi.org/10.1049/ip-sen:19982197

[29] Shirley Gregor. 2006. The Nature of Theory in Information Systems. MIS Quarterly 30, 3 (September 2006), 611–642. https://doi.org/10.2307/25148742

[30] James Hendler. 2014. Data Integration for Heterogenous Datasets. Big Data 2, 4 (December 2014), 205–215. https://doi.org/10.1089/BIG.2014.0068

[31] Daphne van Hesteren, Laura van Knippenberg, Raymonde Weyzen, Esther Huyer, and Gianfranco Cecconi. 2022. *Open Data Maturity Report 2021*. Publications Office of the European Union. Retrieved April 13, 2022 from https://data.europa.eu/doi/10.2830/394148

[32] Alan R. Hevner, Salvatore T. March, Jinsoo Park, and Sudha Ram. 2004. Design Science in Information Systems Research. MIS Quarterly 28, 1 (March 2004), 75–105. https://doi.org/10.2307/25148625

[33] Anne Immonen, Marko Palviainen, and Eila Ovaska. 2014. Requirements of an Open Data Based Business Ecosystem. IEEE Access 2, (February 2014), 88–103. https://doi.org/10.1109/ACCESS.2014.2302872

[34] Hannu Jaakkola, Timo Mäkinen, and Anna Eteläaho. 2014. Open Data: Opportunities and Challenges. In Proceedings of the 15th International Conference on Computer Systems and Technologies (CompSysTech'14), ACM, New York, NY, USA, 25–39. https://doi.org/10.1145/2659532.2659594

[35] Marijn Janssen, Yannis Charalabidis, and Anneke Zuiderwijk. 2012. Benefits, Adoption Barriers and Myths of Open Data and Open Government. Inf. Syst. Manag. 29, 4 (October 2012), 258–268. https://doi.org/10.1080/10580530.2012.716740

[36] Pavel Krasikov, Christine Legner, and Markus Eurich. 2021. Sourcing the Right Open Data: A Design Science Research Approach for the Enterprise Context. In The Next Wave of Sociotechnical Design. Springer, Cham, 313–327. https://doi.org/10.1007/978-3-030-82405-1_31

[37] Pavel Krasikov, Timo Obrecht, Christine Legner, and Markus Eurich. 2020. Is Open Data Ready for Use by Enterprises? Learnings from Corporate Registers. In Proceedings of the 9th International Conference on Data Science, Technology and Applications (DATA'20), SciTePress, 109–120. https://doi.org/10.5220/0009875801090120

[38] Pavel Krasikov, Timo Obrecht, Christine Legner, and Markus Eurich. 2021. Open Data in the Enterprise Context: Assessing Open Corporate Data's Readiness for Use. In Data Management Technologies and Applications, Slimane Hammoudi, Christoph Quix and Jorge Bernardino (eds.). Springer, Cham, 80–100.

[39] Renáta Máchová and Martin Lněnička. 2017. Evaluating the Quality of Open Data Portals on the National Level. J. Theor. Appl. Electron. Commer. Res. 12, 1 (January 2017), 21–41. https://doi.org/10.4067/S0718-18762017000100003

[40] Auriane Marmier and Tobias Mettler. 2020. Different Shades of Perception: How Do Public Managers Comprehend the Re-use Potential of Open Government Data? In Proceedings of the 41st International Conference on Information Systems (ICIS'20).

[41] Sébastien Martin, Muriel Foulonneau, Slim Turki, and Madjid Ihadjadene. 2013. Risk Analysis to Overcome Barriers to Open Data. Electron. J. E-Gov. 11, 1 (December 2013), 348–359.

[42] Xavi Masip-Bruin, Guang-Jie Ren, Rene Serral-Gracia, and Marcelo Yannuzzi. 2013. Unlocking the Value of Open Data with a Process-Based Information Platform. In Proceedings of the 15th Conference on Business Informatics (CBI'13), IEEE, Vienna, Austria, 331–337. https://doi.org/10.1109/CBI.2013.54

[43] Aparna Nayak, Bojan Bozic, and Luca Longo. 2021. (Linked) Data Quality Assessment: An Ontological Approach. In Proceedings of the 15th International Rule Challenge, 7th Industry Track, and 5th Doctoral Consortium (RuleML+RR'21), Technological University Dublin. https://doi.org/10.21427/KRPN-NH58

[44] Sebastian Neumaier, Jürgen Umbrich, and Axel Polleres. 2016. Automated Quality Assessment of Metadata across Open Data Portals. J. Data and Information Quality 8, 1 (October 2016), 1–29. https://doi.org/10.1145/2964909

[45] Open Government Working Group. 2007. The 8 Principles of Open Government Data. Retrieved July 23, 2019 from https://opengovdata.org/

[46] Open Knowledge Foundation. 2005. The Open Definition: Defining Open in Open Data, Open Content and Open Knowledge. Retrieved May 27, 2021 from https://opendefinition.org/

[47] Opendatasoft. 2022. A Comprehensive List of 2600+ Open Data Portals in the World. Open Data Inception. Retrieved April 21, 2022 from https://opendatainception.io/

[48] Edobor Osagie, Mohammad Waqar, Samuel Adebayo, Arkadiusz Stasiewicz, Lukasz Porwol, and Adegboyega Ojo. 2017. Usability Evaluation of an Open Data Platform. In Proceedings of the 18th Annual International Conference on Digital Government Research (dg.o'17), ACM, New York, NY, USA, 495–504. https://doi.org/10.1145/3085228.3085315

[49] Heiko Paulheim. 2016. Knowledge Graph Refinement. Semantic Web 8, 3 (December 2016), 489–508. https://doi.org/10.3233/SW-160218

[50] Ken Peffers, Tuure Tuunanen, Marcus A. Rothenberger, and Samir Chatterjee. 2007. A Design Science Research Methodology for Information Systems Research. J. Manag. Inf. Syst. 24, 3 (December 2007), 45–77. https://doi.org/10.2753/MIS0742-1222240302

[51] Leo L. Pipino, Yang W. Lee, and Richard Y. Wang. 2002. Data Quality Assessment. Commun. ACM 45, 4 (April 2002), 211–218. https://doi.org/10.1145/505248.506010

[52] Nicolas Prat, Isabelle Comyn-Wattiau, and Jacky Akoka. 2015. A Taxonomy of Evaluation Methods for Information Systems Artifacts. J. Manag. Inf. Syst. 32, 3 (July 2015), 229–267. https://doi.org/10.1080/07421222.2015.1099390

[53] Konrad Johannes Reiche, Edzard Höfig, and Ina Schieferdecker. 2014. Assessment and Visualization of Metadata Quality for Open Government

Data. In Proceedings of the 2014 Conference for E-Democracy and Open Governement (CeDEM'14), Donau-Universität, Krems, Austria, 335–346.

[54] Guang-Jie Ren and Susanne Glissmann. 2012. Identifying Information Assets for Open Data. In Proceedings of the 14th International Conference on Commerce and Enterprise Computing (CEC'12), IEEE, Hangzhou, China, 94–100. https://doi.org/10.1109/CEC.2012.23

[55] Erna Ruijer, Stephan Grimmelikhuijsen, Jochem van den Berg, and Albert Meijer. 2018. Open Data Work: Understanding Open Data Usage from a Practice Lens. Int. Rev. Adm. Sci. 86, 1 (May 2018), 3–19. https://doi.org/10.1177/0020852317753068

[56] Kurt Sandkuhl and Ulf Seigerroth. 2019. Method Engineering in Information Systems Analysis and Design. Softw Syst Model 18, 3 (June 2019), 1833–1857. https://doi.org/10.1007/s10270-018-0692-3

[57] David Schatsky, Jonathan Camhi, and Craig Muraskin. 2019. *Data Ecosystems: How Third-Party Information Can Enhance Data Analytics*. Deloitte. Retrieved February 19, 2020 from https://www2.deloitte.com/content/dam/insights/us/articles/4603_Data-ecosystems/DI_Data-ecosystems.pdf

[58] Maung K. Sein, Ola Henfridsson, Sandeep Purao, Matti Rossi, and Rikard Lindgren. 2011. Action Design Research. MIS Q. 35, 1 (March 2011), 37–56. https://doi.org/10.2307/23043488

[59] Mattias Strand and Anna Syberfeldt. 2020. Using External Data in a BI Solution to Optimise Waste Management. J. Decis. 29, 1 (September 2020), 53–68. https://doi.org/10.1080/12460125.2020.1732174

[60] Milena Stróżyna, Gerd Eiden, Witold Abramowicz, Dominik Filipiak, Jacek Małyszko, and Krzysztof Węcel. 2018. A Framework for the Quality-based Selection and Retrieval of Open Data. Electron. Mark. 28, 2 (May 2018), 219–233. https://doi.org/10.1007/s12525-017-0277-y

[61] Antonio Vetrò, Lorenzo Canova, Marco Torchiano, Camilo Orozco Minotas, Raimondo Iemma, and Federico Morando. 2016. Open Data Quality Measurement Framework: Definition and Application to Open Government Data. Gov. Inf. Q. 33, 2 (April 2016), 325–337. https://doi.org/10.1016/j.giq.2016.02.001

[62] Richard Y. Wang and Diane M. Strong. 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. J. Manag. Inf. Syst. 12, 4 (1996), 5–33. https://doi.org/10.1080/07421222.1996.11518099

[63] Frederika Welle Donker and Bastiaan Van Loenen. 2017. How to Assess the Success of the Open Data Ecosystem? Int. J. Digit. Earth 10, 3 (March 2017), 284–306. https://doi.org/10.1080/17538947.2016.1224938

[64] YData Labs. 2023. Pandas profiling overview. Retrieved January 27, 2023 from https://pandas-profiling.ydata.ai/docs/master/index.html

[65] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. 2016. Quality Assessment for Linked Data: A Survey. Semantic Web 7, 1 (2016), 63–93. https://doi.org/10.3233/SW-150175

[66] Ruojing Zhang, Marta Indulska, and Shazia Sadiq. 2019. Discovering Data Quality Problems: The Case of Repurposed Data. Bus. Inf. Syst. Eng. 61, 5 (October 2019), 575–593.

[67] Shichao Zhang, Chengqi Zhang, and Qiang Yang. 2003. Data Preparation for Data Mining. Appl. Artif. Intell. 17, 5–6 (May 2003), 375–381. https://doi.org/10.1080/713827180

[68] Anneke Zuiderwijk, Marijn Janssen, Sunil Choenni, Ronald Meijer, and Roexsana Sheikh Alibaks. 2012. Socio-Technical Impediments of Open Data. J. E-Gov. 10, 2 (January 2012), 156–172.

[69] Anneke Zuiderwijk, Marijn Janssen, and Marijn Janssen. 2014. Barriers and Development Directions for the Publication and Usage of Open Data. In Open Government. Springer, New York, NY, USA, 115–135. https://doi.org/10.1007/978-1-4614-9563-5_8

[70] Anneke Zuiderwijk, Marijn Janssen, Kostas Poulis, and Geerten van de Kaa. 2015. Open Data for Competitive Advantage: Insights from Open Data Use by Companies. In Proceedings of the 16th Annual International Conference on Digital Government Research (dg.o'15), ACM, Phoenix, AZ, USA, 79–88. https://doi.org/10.1145/2757401.2757411

[71] Open Data Barometer. Retrieved April 23, 2022 from https://opendatabarometer.org/?_year=2017&indicator=ODB

[72] Company Register - Global Open Data Index. Retrieved April 23, 2022 from https://index.okfn.org/dataset/companies/

[73] GLEIF - Registration Authorities List. Retrieved April 28, 2022 from https://www.gleif.org/en/about-lei/code-lists/gleif-registration-authorities-list