Data Source Selection Support in the Big Data Integration Process - Towards a Taxonomy

Felix Kruse
Carl von Ossietzky Universität Oldenburg, Department for Business Informatics (VLBA), Oldenburg, Germany

Christoph Schröer
Volkswagen AG, Corporate Foresight, Wolfsburg, Germany

Jorge Marx Gómez
Carl von Ossietzky Universität Oldenburg, Department for Business Informatics (VLBA), Oldenburg, Germany

Data Source Selection Support in the Big Data Integration Process - Towards a Taxonomy

Felix Kruse[1], Christoph Schröer[2] and Jorge Marx Gómez[1]

[1] Carl von Ossietzky Universität Oldenburg, Department for Business Informatics (VLBA), Oldenburg, Germany
{felix.kruse,jorge.marx.gomez}@uni-oldenburg.de
[2] Volkswagen AG, Corporate Foresight, Wolfsburg, Germany
{christoph.schroeer}@volkswagen.de

Abstract. Selecting data sources is a crucial step in providing a useful information base to support decision-makers. While any data source can represent a potential added value in decision making, it's integration always implies a representative effort. For decision-makers, data sources must contain relevant information in an appropriate scope. The data scientist must assess whether the integration of the data sources is technically possible and how much effort is required. Therefore, a taxonomy was developed to identify the relevant data sources for the decision-maker and minimize the data integration effort. The taxonomy was developed and evaluated with real data sources and six companies from different industries. The final taxonomy consists of sixteen dimensions that support the data scientist and decision-maker in selecting data sources for the big data integration process. An efficient and effective big data integration process can be carried out with a minimum of data sources to be integrated.

Keywords: Data Source Selection, Big Data Integration, Taxonomy, Record Linkage, Data Science

1 Introduction

More and more information about real-world entities is digitized and stored in databases. This information can be company-related information such as new product releases, company acquisitions, patent applications, or person-related information such as their employer, published papers, or which competences they have. Many of this information is available in various internal and external databases. These data sources with complementary, additional, or different valuable information are rarely combined, which is why they can be called data silos [1]. There is often a lack of information about the existence and a lack of transparency about the content of the data sources [1]. The reduction of data silos leads to an increase in information value when several data sources are combined [2]. Decision-makers in companies and research need this added information value from combined data sources to make a decision that results in a successful action. This sequence can be described by the big data information value chain [3]. It describes the sequence of (1) data, to (2) information, to (3) knowledge, which is used in a (4) decision, and results in an (5) action. It is crucial to select the data

16th International Conference on Wirtschaftsinformatik,
March 2021, Essen, Germany

sources with the required information that is relevant for the decision-making. Since the required information can be located in different data sources, they must first be integrated. The data integration aims to enable uniform access to data, which are located in several independent data sources [2]. The Big Data Integration (BDI) technical challenges such as semantic, syntactic, and technical heterogeneity between data sources must be overcome to enable data integration [2, 4]. Fig. 1 shows the BDI process extended by the process step of data source *selection*. First, the relevant external and internal data sources must be selected. Then the process steps *schema matching*, *record linkage*, and *data fusion* must be completed to finally obtain the integrated data source [2, 5].

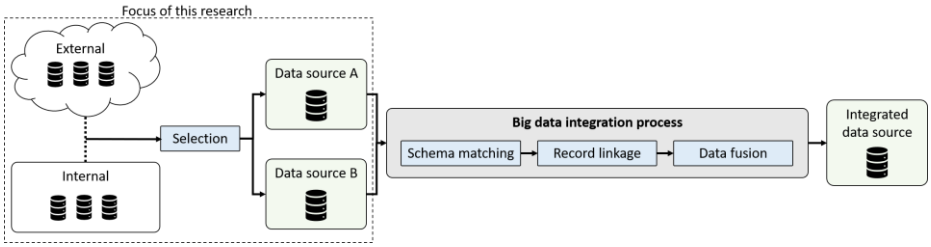

Figure 1. Extended big data integration process to include data source selection [6]

The integrated database is used to develop a data product that supports business decisions. The data product can be a descriptive, predictive, or prescriptive analysis result. The Cross-Industry Standard Process for Data Mining (CRISP-DM) is widely used to develop a data product [6, 7]. In the first CRISP-DM phases business and data understanding, the data product's goal is defined, and the data sources are selected. These tasks are often performed by data scientists and decision-makers from the respective application departments, such as marketing or sales.

With the available number of data sources, there is often the problem that neither the data scientist nor the decision-maker is aware of all of them. The decision-maker cannot judge which data source contains the relevant information. The data scientist cannot estimate the technical effort required to receive and integrate the data sources. It is generally difficult to track the number of data sources, compare them with each other from the point of view of the data scientist and the decision-maker, and ultimately select the most suitable ones. The selection of data sources is a crucial task for carrying out an efficient and effective BDI process [4]. The goal should be to integrate as few data sources as possible to obtain the most appropriate information base. Because the more data sources have to be integrated, the more complex the BDI process is. At this point, the following research question arises, which is to be answered in this paper: "*How can the data source selection be supported in the big data integration process?"* To answer the research question, we used the qualitative research method to develop a taxonomy by Nickerson et al. [8]. Our research approach is inductive, since we obtain generalizing insights from concrete data sources and domain knowledge. The paper is structured as follows. In section 0 the theoretical background is presented. In section 2 the development and evaluation of the taxonomy is described. The final taxonomy is

described in section 3. Afterwards, the use of the taxonomy is described in section 4. Finally, a summary and future research directions are presented in section 5.

Background

Many research papers exist along the BDI process. The literature review of Kruse et al. [9] describes the current state of research in the field of entity linking and record linkage. Entity linking tries to extract relevant entities such as persons or companies from unstructured texts. Record linkage identifies records that refer to the same real-world entity, such as a person or a company. Furthermore, there are record linkage systems like Magellan[1] or the framework BigGorilla[2], which are technically supporting the complete BDI process [10–12]. All these papers focus only on the three process steps *schema matching*, *record linkage*, and *data fusion*.

We conducted a literature search for the relevant topics *data source selection* and *creation of taxonomies to describe data sources*. The relevant research papers from both areas are presented below.

Data source selection research: For the data source selection in the context of big data, papers such as that of Safhi et al. [13] exist. This paper develops an algorithm to identify the subset of relevant and reliable sources with the lowest cost from an existing set of data sources [13]. A prerequisite for the procedure is that all data sources are available and accessible to calculate the developed metrics. Safhi et al. [13] summarize the problem of data source selection as a compromise between the contribution of the source, its quality, and the associated costs.

Assaf et al. [14] developed a framework for assessing the quality of Linked Open Data. They developed a tool that profiles the data sources and evaluates them based on objectively measurable indicators. Nevertheless, the reference to the BDI process is missing in this paper.

Lin et al. [4] develop an algorithm for the evaluation of data quality. The algorithm calculates the number of expected correct values per attribute for a data source (truth discovery). With this single criterion, the data sources with the truest attributes can be selected [4]. The procedure requires full access to the data source to execute the algorithms. It also targets only the truth content criterion and helps to select the data sources with the highest truth content. A reference to the BDI process is missing.

Dong et al. [15] aim to support the selection of data sources before the BDI process starts so that the quality of the data and the data integration effort can be balanced. However, first, the authors focus on the last step of the BDI process, the data fusion, in which the conflicts of the already integrated data sources must be solved [15]. Building on this, Rekatsinas et al. [16] go further into detail by extending the approach of Dong et al. [15] for changing data sources.

Data source taxonomy research: There also exists research work to classify data sources with the help of a taxonomy. In the paper of Zrenner et al. [17] a data source taxonomy for the visibility of the supply chain network structure is developed. The

[1] https://sites.google.com/site/anhaidgroup/projects/magellan
[2] https://www.biggorilla.org/

taxonomy goal is to increase the knowledge of practitioners and researchers about data sources for supply chain network structures. According to Zrenner et al. [17], the taxonomy should support the initial data source selection. However, the taxonomy is limited to supply chain data sources and does not reference the BDI process.

Li et al. [18] present a rule-based taxonomy of dirty data. The taxonomy is designed to support companies in better monitoring, analyzing, and cleansing dirty data. The authors present a method to solve the problem of dirty data selection, since often not all data cleansing procedures can be performed due to limited computing capacity [18]. This taxonomy focuses on the support of the general data preparation and not the BDI process.

Roeder et al. [19] present a taxonomy to classify the heterogeneity of data sources and help researchers and professionals explore data sources. The authors consider the 5V definition of big data (volume, velocity, variety, value, and veracity) when developing the taxonomy. However, the value and the veracity of the data sources are not taken into account. From the author's perspective, the added value of the data is challenging to measure objectively [19]. The taxonomy is evaluated by applying it to five other data sources. The paper lacks an evaluation with practitioners or researchers who were not involved in taxonomy development [19].

The presented papers show that research in the areas is conducted separately. No paper considers the creation of a taxonomy for data source selection. Also, no paper in both areas consider the BDI process. This research gap is investigated in our paper.

2 Development process of the Data Source Taxonomy

The classification of objects of a domain into a taxonomy is a problem in many disciplines, such as information systems research. A taxonomy supports structuring and organizing knowledge of a defined domain. A taxonomy enables researchers to describe and investigate the relationships between the concepts captured in the taxonomy. Taxonomies as structure-giving artifacts play a key role in the exploration of new fields of research in Information Systems (IS) [8, 20]. A taxonomy T is defined as a set of n dimensions. Each of these dimensions contains mutually exclusive and overall complete characteristics [8]. Nickerson et al. [8] define that only one characteristic from each dimension may be assigned to an object [8, 19]. The taxonomy we have created allows for multiple selections of characteristics to increase the taxonomy's usefulness.

For the development of a taxonomy Nickerson et al. [8] have developed a widely used process. The process supports researchers in developing a taxonomy [8]. The process is shown in Fig. 2 and is described in the next section.

2.1 Development of the taxonomy

In the first process step, *determine meta-characteristics*, the goal of the taxonomy should be defined. Based on the defined goal, the dimensions and characteristics can be determined in a targeted manner. Nickerson et al. [8] recommend deriving the goal from the potential users and the related use cases of the taxonomy [8].

Meta-characteristic: The taxonomy is intended to support a data scientist and a decision-maker in selecting data sources in the BDI process. The content of the data source should be described to estimate the added value of the information. Also, technical characteristics should be described in order to be able to estimate the possibilities and the effort required for data integration.

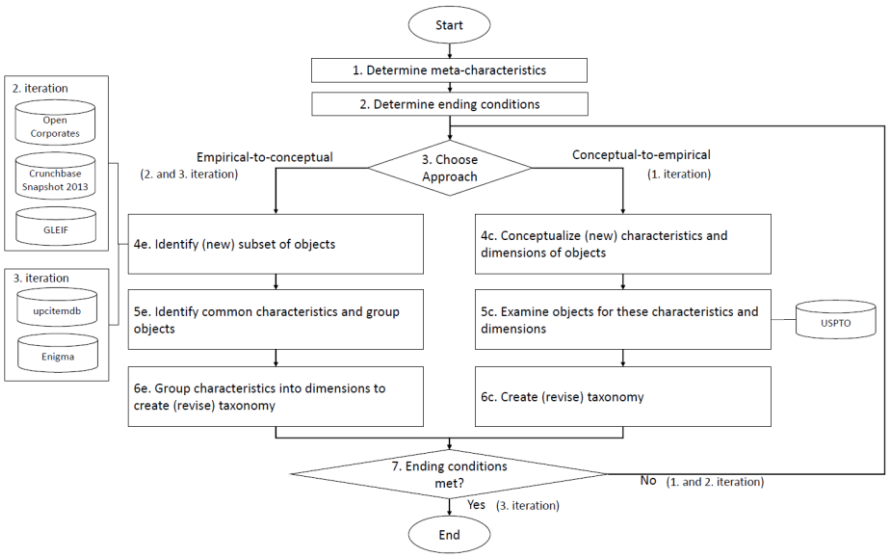

Figure 2. Taxonomy development method [21]

In the second process step, *determine ending conditions*, objective and subjective ending conditions for the process are defined. These ending conditions are necessary to stop the iterative process [8]. This paper adopts the eight objective and five subjective ending conditions proposed by Nickerson et al. [8] (see table 1). After each iteration, the ending conditions are checked (process step 7 *ending conditions met?*). The process stops when all conditions are met.

The iterative process begins in step three *choose approach*. In this process step, a decision is made between the empirical-to-conceptual and the conceptual-to-empirical approach. The choice of the approach is determined by the domain knowledge and the available objects. In our taxonomy development process, the data sources represent the objects.

The conceptual-to-empirical approach is recommended if a few data sources, but significant domain knowledge is available. The empirical-to-conceptual approach is recommended if little domain knowledge but many data sources are available [8, 19].

For the first iteration, we could decide between both approaches since the authors have domain knowledge and six relevant data sources. We decided to use the conceptual-to-empirical approach in the first iteration.

For this purpose, the dimensions and characteristics are first derived from existing theories (process step 4c *conceptualize (new) characteristics and dimensions of objects*).

For the initial creation of the taxonomy, we refer to the paper by Zrenner et al. [17], who developed a data source taxonomy for the field of supply chain management without reference to the BDI process. Their created taxonomy was first generalized so that the taxonomy can be used for any application area.

Table 1. Final conditions fulfilled per iteration of the taxonomy development [8]

Iteration			*Ending Condition*
1	***2***	***3***	***Objective condition***
		x	All relevant objects have been examined
x	x	x	No merge or split of objects
	x	x	Each characteristic of each dimension was selected by one object
		x	No new dimension or characteristic was added
		x	No dimension was merged or split
x	x	x	Every dimension is unique
x	x	x	Every characteristic is unique within its dimension
x	x	x	Each combination of characteristics is unique and is not repeated
			Subjective condition
	x	x	Concise: Meaningful without being unwieldy or overwhelming
	x	x	Robust: Significant and informative characteristics
		x	Comprehensive: All objects or a sample of objects can be classified
x	x	x	Extendible: Dimensions and characteristics can be easily added
x	x	x	Explanatory: The dimensions and characteristics explain the objects

The information quality of the data sources is crucial for the selection for integration [2]. Therefore, we take into account the widely used 15 information quality (IQ) dimensions of Wang et al. [21]. The IQ dimensions are divided into the superordinate categories, (1) intrinsic data quality, (2) contextual data quality, (3) representational data quality, and (4) accessibility data quality. The criteria give an overview of relevant evaluation dimensions of data sources [21], which are valid until today.

Furthermore, the taxonomy of Roeder et al. [19] is included in the development in order to consider the 5V (volume, velocity, variety, value, and veracity) of big data in the development of the taxonomy. In contrast to Roeder et al. [19], we objectively describe the value of a data source with our taxonomy. We will also try to describe the veracity property of big data objectively to a certain extent. The trustworthiness of a data source is a difficult criterion to measure. We think that this can only be reliably estimated by working with the data and independently checking the data and that only a rough tendency can be made when selecting data sources that have not yet been worked with in detail. In the next process step, *5c examine objects for these characteristics and dimensions*, the data source of the United States Patent and Trademark Office (USPTO)[3] was applied to the taxonomy. Then the process step *6c create(revise) taxonomy* was performed and it was determined that not all ending conditions were met (see table 2). For the second iteration, we have chosen the

[3] https://developer.uspto.gov/product/patent-grant-bibliographic-dataxml

empirical-to-conceptual method. First, in step *4c identify (new) subset of objects* we used the data sources OpenCorporates[4], Crunchbase Open Data Map[5], Crunchbase 2013 Snapshot[6] and Level-1 dataset from the Global Legal Entity Identifier (GLEIF) Foundation[7].

With these data sources, the taxonomy was further developed in the process steps *5e identify common characteristics and group objects* and *6e group characteristics into dimensions to create (revise) taxonomy*. Even after the second iteration, not all ending conditions were fulfilled. The third iteration was performed with the empirical-to-conceptual method. The data sources upcitemdb[8] and a dataset of the Enigma platform[9] were used for further development. After the third iteration, all final conditions were met and the taxonomy development process was finished.

2.2 Evaluation of the taxonomy

The iterative taxonomy development process and the subjective and objective ending conditions lead to an ex-ante evaluation [20]. The main goal of a taxonomy is to be useful and suitable for the defined use case (meta-characteristics). Since usefulness is a criterion that is difficult to measure, the taxonomy should be presented to and used by the addressed target group [8, 20]. The target group consists of data scientists and decision-makers in our case.

Szopinski et al. [20] presents a framework for the ex-post evaluation of a taxonomy, which is applied in this paper. The framework is divided into the following three sections.

1. Who, subject of evaluation? The evaluation of a taxonomy can be performed by persons who have already been involved in the development of the taxonomy or who are new to the research project for evaluation. These persons can be researchers or practitioners [20]. For the evaluation, we were able to draw on researchers from the University of Oldenburg and Goettingen (see table 2), who are experienced in both the development of a taxonomy and the selection of data sources. Furthermore, we were able to win over data scientists and decision-makers from different sectors like automotive OEM, software development, photo and online print service, energy utility, energy sales, and financial services to evaluate the taxonomy. With this evaluation partners, the target group of the taxonomy is covered.

2. What, object of evaluation? The evaluation can be performed directly with data sources (objects) or indirectly with papers reporting on data sources (research on the objects). Therefore data sources can be used, which have already been used for the development of the taxonomy or completely new data sources [20]. The evaluation in

[4] https://opencorporates.com/

[5] Powered by Crunchbase: https://data.crunchbase.com/docs/open-data-map

[6] Crunchbase 2013 Snapshot ©, Creative Commons Attribution License [CC-BY], https://data.crunchbase.com/docs/2013-snapshot

[7] https://www.gleif.org/de/lei-data/gleif-concatenated-file/download-the-concatenated-file

[8] https://www.upcitemdb.com/

[9] At the time of access still freely available: https://public.enigma.com/browse/collection/stock-exchanges-company-listings/50a2457d-6407-4581-8f14-5d37a9410fa9

this paper was done in one case (ID 1) with data sources that have already been used for the development of the taxonomy (see table 2). The evaluation runs with the IDs 3, 6, 7, and 9 based on the participating persons' expertise. The remaining evaluation runs were performed with new data sources.

Table 2. Overview of the evaluation of the taxonomy

Who			*What*	*How*		
ID	***Sector***	***Role***	***Object (data source)***	***Focus group***	***Expert interview***	***Illustrative interview***
1	University Oldenburg	Researcher	OpenCorporates, Crunchbase			x
2	Automotive OEM	Decision-maker	Internal data sources	x		x
3	University Goettingen	Researcher	About real-world		x	
4	Software Development	Data Scientist	Covid, Natural Earth, Wiki		x	x
5	Photo and online print service	Data Scientist	Weather data source	x		x
6	Energy Utility	Decision-maker	About real-world		x	
7	Energy Sales	Decision-maker	About real-world		x	
8	University Oldenburg	Researcher	UTKFace, IMDB-WIKI	x		x
9	Energy Utility	Data Scientist	About real-world		x	
10	Financial Services	Data Scientist	Internal data sources	x		x

3. How, method of evaluation? The evaluation can be carried out with different methods, which are described in the paper by Szopinski et al. [20]. We have chosen the methods expert interview, focus group, and illustrative scenario (with real data sources). The expert interview was used when one person out of the target group was available for evaluation (see table 2 ID's 3, 4, 6, 7, and 9). The focus group was used if more than one person out of the target group was available (see table 2 IDs 2, 5, 8, and 10). In both methods, we first introduced the taxonomy to the persons and then asked the following open questions recommended by [20]: (1) Is the taxonomy understandable and complete? (2) Have all relevant objects been considered in the taxonomy? (3) Which dimensions or characteristics should be changed, added or deleted?

In the evaluation method illustrative scenario (see table 2 ID's 1, 2, 4, 5, 8, 10) the taxonomy was applied by the respective evaluation partners to data sources such as

weather data (ID 5), Covid[10], Natural Earth[11], Wiki[12] data (ID 4), UTKFace[13], IMDB-Wiki[14] (ID 8) or to internal company data (ID 2 and 10).

The feedback from the evaluation runs has led to adjustments to the taxonomy, so that it has been iteratively developed further (see table 1). The subjective and objective ending conditions from table 1 were used again to determine the end of the evaluation runs. After ten evaluation runs, all ending conditions were met, so that the final taxonomy was developed to assist in selecting data sources.

3 The final taxonomy

This section describes the final taxonomy (see Fig. 3). The taxonomy is intended to be used by data scientists and decision-makers who select data sources for the BDI process. We think that the selection of data sources depends strongly on the data product. The data source taxonomy should capture as objective criteria as possible to support an optimal decision for individual data products.

Dimension	Characteristics					
D1: Accessibility	Internal	External (open)	External (closed)			
D2: Licensing	Specification Open Source License	Provider own license	Not available			
D3: Use after license expiry	Data can be further used	Data may no longer be used and must be deleted	Not available			
D4: Price model	Quantity-controlled	Time-controlled	One time costs	Free of charge	Data owner	
D5: Interface	API	GUI	Manual download	Data medium		
D6: Data structure	Strucutred	Semistructured	Unstructured			
D7: Reported point in time or period	Period of time	Point in time	Not available			
D8: Update	Real-time	Regular interval	Not available			
D9: Language	Source language(s)	Translated into language(s)	Not available			
D10: Scope of the data source	Complete	Self-selected extract	Provided extract			
D11: Preprocessing of the data	Schema created	Metadata generated	Metadata from the data provider	Keep original data format		
D12: Current data status	Specification of the data status (date or version)					
D13: Real-world entity	Company	Person	Product	Patent	Geographical location	[...]
D13a: Number of records	Specification of the number	Not available				
D13b: Data volume	Specification of the volume	Not available				
D13c: Number of describing attributes	Specification of the number of descriptive attributes	Not available				
D14: Total data volume	Specification of the total volume	Not available				
D15: Number of tables or files	Specifying the number of tables or files	Not available				
D16: Added Information Value	Balance sheet data	Mergers and acquisitions information	Product information	Financing and stock exchange data	Corporate structures	Pantet applications, patent grants	[...]

Figure 3. Final data source taxonomy to support the data source selection

D1: Accessibility The dimension accessibility was taken over from the taxonomy of [17] and at the same time addresses the IQ dimension 7 accessibility [21]. The dimension has the characteristics C_{1j} = {*Internal*, *external(open)*, *external(closed)*}. A distinction is made between internal and external data sources from the perspective of the user of the taxonomy. For external data sources, there is also a distinction between

[10] https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset
[11] http://www.naturalearthdata.com/downloads/10m-cultural-vectors/
[12] https://www.kaggle.com/juanumusic/countries-iso-codes
[13] https://susanqq.github.io/UTKFace/
[14] https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/

whether login data is required for access (*external(closed)*) or whether it is accessible without barriers (*external(open)*).

D2: Licensing The dimension License was created during the three development iterations. It has the characteristics C_{2j} = {*Specification open-source license, provider own license, not available*}. Under the characteristic *specification open source license* the existing open source license should be specified such as MIT or BSD-3-Clause. Commercial data source providers often conclude individual license agreements. Then the characteristic *provider own license* should be selected. If nothing is known about the license, select *not available*.

D3: Use after license expiry The dimension use after license expiry was created by the evaluation with the practitioners. It has the characteristics C_{3j} = {*Data can be further used, data may no longer be used and must be deleted, not available*}. The dimension is intended to describe the data sources in terms of how to deal with data after the license expires.

D4: Price model The dimension price model is taken from the taxonomy of Zrenner et al. [17]. It has the characteristics C_{4j} = {*Quantity-controlled, time-controlled, one time costs, free of charge, data owner*}. This dimension should describe the pricing model of the data source. In this dimension, multiple selections are possible, since, for example, a combination of a *quantity-controlled* and *time-controlled* pricing model is possible. The base account of OpenCorporates with 20000 requests per month is an example for such pricing models. If an internal data source is classified, the characteristic *data owner* should be selected.

D5: Interface The Interface dimension was created during the development-iterations and is intended to describe the user's access options to the data source. The characteristics C_{5j} = {*API, GUI, manual download, data medium*} serve this purpose. Multiple selections are possible. When selecting the characteristics, it is best to specify which data formats such as XML, JSON, or CSV are offered. The characteristic *data medium* is selected if the data source is provided e.g., via a hard disk or USB stick.

D6: Data structure The dimension data structure is described by the characteristics C_{6j} = {*Schema(structured), schemeless(semi-structured or unstructured)*} whether the data source is structured, semi-structured or unstructured. The dimension was created during the development process.

D7: Reported point in time or period The dimension reported point in time or period was created during the evaluation. The characteristics C_{7j} = {*Period of time, point in time, not available*} are intended to describe the point in time or period covered by the data in the data source. For example, patent data from the USPTO exists since 1976, whereas an overview of AI start-ups only exists for the point in time July 2019.

D8: Update frequency The dimension update frequency describes the update of the current data source with the characteristics C_{8j} = {*Real-time, regular interval, not available}.* The data source can be updated continuously in real-time or at a certain frequency, which should be specified if possible. If nothing is known about updating the data source, *not available* is selected. This dimension was created during the development process and address IQ Dimension 9 (timeliness) of Wang et al. [21].

D9: Language The Language dimension describes the language used in the data source. The dimension was created during the development process and was extended

during the evaluation. The dimension has the characteristics C_{9j} = {*Source language(s)*, *translated into language(s)*, *not available*}. The languages that appear in the data source should be specified, such as German or English. During the evaluation, a data source was classified, which was translated into a common language by the data provider, for which the characteristic *translated into language(s)* was included. If the data source does not contain a language, but consists, for example, only of numerical values, *not available* is selected. This characteristic also arose during the evaluation of a practice partner who classified a sensor data source.

D10: Scope of the data source This dimension should describe the scope of the data source for classification into this taxonomy. The characteristics C_{10j} = {*Complete*, *self-selected extract of data*, *provided extract of the data*} should be used for this purpose. If the data source is not complete, it is necessary to specify the user's criteria to make a selection or by the data provider. The dimension has been defined during the development-iterations. For example, Crunchbase provides an extract of the data from 2013.

D11: Preprocessing of the data With the dimension preprocessing of the data, it is to be described whether the data source has already been preprocessed and on this basis the classification with the taxonomy is carried out. The characteristics C_{11j} = {*Schema created or metadata generated (structured)*, *structured metadata from the data provider*, *keep original data format*} are to be used for this. For example, JSON files will be preprocessed and converted to a structured format to get a first overview of the data source. Data providers of unstructured data, such as news data, often provide structured metadata for them. If a structured data structure already exists, the data structure is often not changed. This dimension was created during the evaluation process with the practice partners.

D12: Current data status The dimension current data status has one characteristic C_{12j} = {*Specification of the data status (date or version)*} with which the current content status of the data source is to be indicated. The dimension was created during the development process.

D13: Real-world entity This dimension is used to describe which real-world entities are represented in the data source. In the taxonomy in Fig. 3, the last cell (*[...]*) indicates that the characteristics should and may be supplemented by further entities. From the development and evaluation process the characteristics C_{13j} = {*Company*, *person*, *product*, *patent*, *geographical location*} have emerged. This dimension is crucial for the BDI process since it is possible to identify whether and via which entity the data sources could potentially be connected. The goal of the process step record linkage is to identify records that belong to the same real-world entity [5].

D13a: Number of records; **D13b: Data volume**; **13c Number of describing attributes** The dimensions 13a, 13b, and 13c should be filled in for each real-world entity, if possible. The specification of how many unique data records, how large the data volume, and how many describing attributes exist for each real-world entity should help evaluate the value and veracity of the data source. The objective, quantifiable criteria allow the assessment of whether the data source is potentially useful or not for the data product. Also, the number of descriptive attributes serves as a first indication for the execution of the BDI process steps schema matching and record linkage. Since

it can be estimated how many attributes a comparison of the data records can be carried out. The unique number of data records and attributes correlated with the data volume can be used to estimate how complete the data source is.

Furthermore, whether the data source offers an appropriate scope (IQ dimension 19) and thus also relevance (IQ dimension 2) for the respective data product [21]. Other taxonomies like the one from Zrenner et al. [17] or Roeder et al. [19] use characteristics like high, medium, low, which are very subjective. This subjective criteria are difficult to use to compare different data sources. Our chosen objective numerical criteria can be used to compare different data sources.

D14: Total data volume This dimension should cover the entire data volume of the data source. If this is not available, the characteristic *not available* is used. This objective criterion also serves to evaluate the appropriate scope of the data source in comparison to other data sources.

D15: Number of tables or files This dimension should describe the number of existing tables or files of the data source. This objective criterion is intended to provide a first assessment of whether the scope is appropriate (IQ dimension 19) and the information can be relevant (IQ dimension 2) [21].

D16: Added information value This dimension was created during the development process and has been further extended during the evaluation process. With this dimension and its characteristics, the practitioners and researchers had the greatest difficulties understanding and applying it during the evaluation. This dimension should serve to objectively capture the big data characteristics value and the IQ-Dimension 2 value-added for the data source. The final taxonomy (see Fig. 3) shows some characteristics. Multiple selections are possible and the characteristics should be expandable if further data sources with new added information values are captured.

On the one hand, the characteristics must not be recorded in too much detail, as the effort to apply the taxonomy could become too high. On the other hand, the characteristics must not be recorded too roughly, so that the added value of the data source is adequately captured. An important requirement for the operationalization of the taxonomy is that the characteristics of this dimension are maintained and extended centrally so that the instances of the characteristics remain disjunctive.

4 Application of the taxonomy in the data source selection

We applied the final taxonomy to the Crunchbase[6], USPTO Patent Grants[3] and AI Startups[15] data sources to demonstrate the applicability and utility (Fig. 4). In the taxonomy meta-characteristic, it has been defined that the taxonomy users are data scientists and decision-makers. The users should be supported in the process steps business understanding and data understanding when designing a data product. Since the crucial data source selection for the data product development takes place in these phases. To demonstrate the utility of the taxonomy, section 5.1 describes the data integration perspective and section 5.2 the decision-maker perspective of the taxonomy.

[15] https://de.appanion.com/startups

Dimension	Crunchbase[6]				USPTO Patent Grants[3]				AI Startups[15]	
D1: Accessibility	External (closed)				External (open)				External (open)	
D2: Licensing	Provider own license				Provider own license				Not available	
D3: Use after license expiry	Not available				Data can be further used				Not available	
D4: Price model	Time-controlled				Free of charge				Free of charge	
D5: Interface	API	GUI	Manual download		API	GUI	Manual download (XML)		Manual download (Image, PDF, PowerPoint)	
D6: Data structure	Strucutred (MySql Dump)				Semistructured (XML)				Semistructured (PDF)	Unstructured (Image, PowerPoint)
D7: Reported point in time or period	Not available				1976 until today				July 2019	
D8: Update	Real-time				Regular interval				Not available	
D9: Language	english				english				deutsch	
D10: Scope of the data source	Provided extract (Crunchbase 2013 Snapshot)				Self-selected extract (One week 20.08 - 27.08.19)				Complete	
D11: Preprocessing of the data	Keep original data format				Schema created				Schema created	
D12: Current data status	2013				August 2019				Not available	
D13: Real-world entity	Company	Person	Product	Geographical location	Company	Person	Patent	Geographical location	Company	Geographical location
D13a: Number of records	118.342	117.318	25.059	7.976	34.409	209.123	114.138	24.682	279	47
D13b: Data volume	166 MB				12 MB	25 MB	14 MB	12 MB	32 KB	
D13c: Number of describing attributes	40				10	8	7	3	4	1
D14: Total data volume	300 MB				200 MB				32 KB	
D15: Number of tables or files	10 tables				9 tables				1 tables	
D16: Added Information Value	Worldwide				Patent applications in america				German ai-startups	
	Financing and stock exchange data		Corporate structures		Patent applications, patent grants				Funding and investor information	
	Relationships between Company, Person, Product				Relationships between Company, Person, Patent				Branch and application focus	
	Graduation of persons		Mergers and acquisitions information		Technology classification				Strategic startup orientation	

Figure 4. Application of the taxonomy on the data sources Crunchbase, USPTO and AI Startups

4.1 Data integration perspective

From the data integration point of view the following questions could be answered for example:

- Q1: Which real-world entity(ies) can be used to link the data sources?
- Q2: How many attributes are available for comparison?
- Q3: How is the data source structured?
- Q4: What is the data volume of the entities?
- Q5: How can the data source be accessed?

(Q1) The dimension D13 provides the information that the three data sources could be integrated via the entity company or geographical location. Integration via the patent entity of the USPTO Patent Grants data source is not possible because no other data source contains this entity.

(Q2) The dimension 13a contains the number of attributes for each real-world entity. This information is used for a first estimation of how successful and sophisticated the integration could be since the attributes that can be compared are identified in the BDI process step schema matching [5]. All Crunchbase entities are stored in a common table consisting of 40 attributes. The USPTO Patent Grants contains ten attributes for the entity company. The AI Startups contains four attributes for the entity company. For the integration of the AI startups with one of the other data sources, the four attributes must be mapped to the 10 or 40 attributes (schema matching). We assume that more attributes improve the quality of the BDI process result, but also increase complexity.

(Q3) The structure of the data source can be read from D6. The BDI process's effort increases if semi-structured or unstructured data sources are available because the BDI process requires structured data.
(Q4) If there is only a part of the data source available (D10) and data integration is to be carried out with this part, dimension 11 is relevant. Dimension 11 documents whether the original data structure has been retained or preprocessed. For example, the original XML structure (D6) of the USPTO Patent Grants was converted into a structured form (D11) with nine tables (D15). The number of data records (D13a) and the data volume (D13b) can be used to estimate the computing capacity required for data integration. If the data sources Crunchbase and AI Startups are to be integrated via the entity company, 33.017.418 (279 x 118.342) data records would have to be compared. At this point, the data scientist can get a first assessment of a suitable blocking algorithm to reduce the number of comparisons in the record linkage process.
(Q5) The dimension D5 provides the information on how to access the data source. The data source AI Startups only offers a manual download of the AI Startup Report in the data formats Image, PDF, and PowerPoint. This dimension is essential for determining the degree of automation of the subsequent operationalization of the data product.

4.2 Decision-maker perspective

We think that the choice of decision-makers data sources depends on the goal of the data product. Therefore, the taxonomy provides objective and comparable criteria that can be individually evaluated and prioritized for each data product.
The taxonomy allows the decision-maker to answer the following questions:

- Q1: Do the data sources contain useful information, added value, and appropriate scope for the data product?
- Q2: Are the data sources sufficiently reliable?
- Q3: Is the data current enough and goes back far enough into the past?
- Q4: What is the licensing model of the data sources?
- Q5: How expensive is the data source?

(Q1) The added value of the information provided by the data source can be taken from the dimension D16. If, for example, AI startups are required for a data product, the AI startup's data source is suitable. Patent information for the data product should also be used. The dimension D16 indicates the decision-maker that the AI startups data source does not provide this information and that the USPTO Patent Grants data source should be used. However, this data source only provides patent grants from America. These value-added information categories can thus be used to select an initial selection of data sources that are suitable for the data product. Via the dimension D10, the decision-maker can see the basis on which the descriptions of D11 - D16 have been collected. With the dimensions D13, D13a, D13b, D13c, decision-makers can also estimate whether the data sources contain a sufficient scope for the data product. If, for example, german AI companies are to be analyzed and the decision-maker knows that about 1000 of such companies, the decision-maker recognizes that the AI Startups data source does not include all german AI companies.

(Q2) The trustworthiness of a data source is a difficult criterion to measure and we think that this can only be reliably estimated by working with the data and independently checking the data. The IQ dimensions believability (1), completeness (10), accuracy (4), and interpretability (5) [21] are covered by the big data characteristic veracity. We think that by specifying the data provider name, the license model (D2), the pricing model (D4), and the dimensions D13a, b, and c, the first estimation of the veracity of the data source can be supported.
(Q3) The up-to-dateness of the data source information can be read from D7 and D8. In D7, it is indicated whether the data source only represents a point in time, like the AI Startup data source, or whether it represents a period, like the USPTO Patent Grants from 1976 until today. In D8, it is indicated whether and how the data source is updated. For example, the USPTO Patent Grants data source is updated weekly. The update frequency of the data source is essential for the operationalization of the data product. Since a decision that has to be made daily, often requires a data source with information that is updated daily. Therefore, the update frequency is a knock-out criterion for the feasibility of a data product.
(Q4) With the dimension licensing (D2) and use after license expiry (D3), the decision-maker gets the information about the license model of the data source. It is equally important to consider the use of the data after the license expires. The reason for this is that any data products developed with this data may no longer be used after the license expires.
(Q5) With the dimension pricing model (D4), the decision-maker can estimate the cost of the data source and put it into a cost-benefit relation when evaluating his data product.

5 Conclusion and further research

The data source selection is a crucial step to develop a useful data product. Therefore, we have extended the BDI process to include the data source selection process step. We have shown that research exists in data source selection, taxonomy development for data sources, and the BDI process. However, these research areas have so far been considered mainly in isolation. With this paper, we try to link the research areas and defined the following research question: *How can data source selection be supported in the big data integration process?* To answer this research question, we developed a data source taxonomy according to the methodical approach of Nickerson et al. [8]. The taxonomy was evaluated according to the evaluation framework of Szopinski et al. [20] with data scientists and decision-makers from two universities and six companies from different sectors. For the development and evaluation of the taxonomy, real data sources such as OpenCorporates, Crunchbase 2013 Snapshot, Upcitemdb, or GLEIF were used. The final taxonomy consists of sixteen dimensions and describes a data source in terms of content and technical criteria to support data scientists and decision-makers in selecting data sources in the BDI process. For example, the taxonomy provides an overview of the added information value of a data source in the form of categories. It

also provides the real-world entities it contains, which can be used to integrate other data sources.
The data source taxonomy developed by us for selection support directly influences theory and practice. Our evaluation shows that companies see taxonomy as support.
Also, decision-makers can use the taxonomy to compare data providers and support a purchase decision based on the completed taxonomies. The taxonomy could be filled out by the data providers to reduce the decision-makers effort. The decision-maker can obtain an overview of the data sources that could potentially be purchased. Our research shows that the big data integration process, defined by [2], should be extended to include the process step data source selection. Our research shows that a taxonomy is suitable to structure and organize the many important aspects of data sources. At the same time, there are some limitations to our work. With the developed taxonomy, we have taken the first step into researching a data source taxonomy. There are many more data source relevant aspects that we have not considered, like security, privacy, compliance, GDPR, data anonymization, or company-specific organizational challenges. All these limitations offer the potential for future research and further development of the data source taxonomy. Also, the taxonomy should be evaluated and further developed in other companies and with other data sources. Especially dimension 16 (added information value) should be researched in more detail.

References

1. Stonebraker, M., Ilyas, I.: Data Integration: The Current Status and the Way Forward. IEEE Data Eng. Bull. 41, 3-9 (2018)
2. Dong, X.L., Srivastava, D.: Big Data Integration. Synthesis Lectures on Data Management 7, 1–198 (2015)
3. Abbasi, A., Sarker, S., Chiang, R.: Big Data Research in Information Systems: Toward an Inclusive Research Agenda. JAIS 17, I–XXXII (2016)
4. Lin, Y., Wang, H., Li, J., Gao, H.: Data Source Selection for Information Integration in Big Data Era (2016)
5. Christen, P.: Data Linkage: The Big Picture. Harvard Data Science Review 1 (2019)
6. Wirth, R.: CRISP-DM: Towards a standard process model for data mining. In: Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, pp. 29–39 (2000)
7. Kruse, F., Dmitriyev, V., Marx Gómez, J.: Building a Connection Between Decision Maker and Data-Driven Decision Process. Archives of Data Science, Series A (Online First) 4, 16 S. online (2018)
8. Nickerson, R.C., Varshney, U., Muntermann, J.: A method for taxonomy development and its application in information systems. European Journal of Information Systems 22, 336–359 (2013)
9. Kruse, F., Hassan, A.P., Awick, J.-P., Marx Gómez, J.: A Qualitative Literature Review on Linkage Techniques for Data Integration. In: Tung Bui (ed.) 53nd Hawaii International Conference on System Sciences, HICSS 2020, Grand

Wailea, Maui, Hawaii, USA, January 7-10, 2020, pp. 1063–1073. ScholarSpace / AIS Electronic Library (AISeL) (2020)
10. Konda, P., Naughton, J., Prasad, S., Krishnan, G., Deep, R., Raghavendra, V., Das, S., C., P.S.G., Doan, A., Ardalan, A., et al.: Magellan: Toward Building Entity Matching Management Systems. Proc. VLDB Endow. 9, 1197–1208 (2016)
11. Konda, P., Subramanian Seshadri, S., Segarra, E., Hueth, B., Doan, A.: Executing Entity Matching End to End: A Case Study. In: Melanie Herschel, Helena Galhardas, Berthold Reinwald, Irini Fundulaki, Carsten Binnig, Zoi Kaoudi (eds.) Advances in Database Technology - 22nd International Conference on Extending Database Technology, EDBT 2019, Lisbon, Portugal, March 26-29, 2019, pp. 489–500. OpenProceedings.org (2019)
12. Govind, Y., Konda, P., Suganthan G C, P., Martinkus, P., Nagarajan, P., Soundararajan, A., Li, H., Mudgal, S., Ballard, J., Zhang, H., et al.: Entity Matching Meets Data Science: A Progress Report from the Magellan Project (2019)
13. Safhi, H.M., Frikh, B., Ouhbi, B.: Data Source Selection in Big Data Context. In: Indrawan-Santiago, M., Pardede, E., Salvadori, I.L., Steinbauer, M., Khalil, I., Anderst-Kotsis, G. (eds.) Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services, pp. 611–616. ACM, New York, NY, USA (2019)
14. Assaf, A., Senart, A., Troncy, R.: Towards An Objective Assessment Framework for Linked Data Quality. International Journal on Semantic Web and Information Systems 12, 111–133 (2016)
15. Dong, X.L., Saha, B., Srivastava, D.: Less is more. Proc. VLDB Endow. 6, 37–48 (2012)
16. Rekatsinas, T., Dong, X.L., Srivastava, D.: Characterizing and selecting fresh data sources. In: Dyreson, C., Li, F., Özsu, M.T. (eds.) Proceedings of the 2014 ACM SIGMOD international conference on Management of data - SIGMOD '14, pp. 919–930. ACM Press, New York, New York, USA (2014)
17. Zrenner, J., Hassan, A.P., Otto, B., Marx Gómez, J.C.: Data source taxonomy for supply network structure visibility. epubli (2017)
18. Li, L., Peng, T., Kennedy, J.: A Rule Based Taxonomy of Dirty Data. GSTF INTERNATIONAL JOURNAL ON COMPUTING 1 (2011)
19. Roeder, J., Muntermann, J., Kneib, T.: Towards a Taxonomy of Data Heterogeneity. In: Gronau, N., Heine, M., Poustcchi, K., Krasnova, H. (eds.) WI2020 Zentrale Tracks, pp. 293–308. GITO Verlag (2020)
20. Szopinski, D., Schoormann, T., Kundisch, D.: Because your taxonomy is worth it: Towards a framework for taxonomy evaluation. In: Proceedings of the Twenty-Seventh European Conference on Information Systems (ECIS) (2019)
21. Wang, R.Y., Strong, D.M.: Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems 12, 5–33 (1996)