BRIDGING THE DATA ANALYTICS – DATA EXCELLENCE GAP: REQUIREMENTS FOR A DATA CATALOGUE FOR EFFICIENT CROSS-ORGANIZATIONAL MOBILITY AND SMART CITY DATA ANALYTICS

Johannes Sautter[1], Jan Kräck[2], Rudolf Fischer[3], Nicolas Fähnrich[4], Kai Erlenhardt[5], Marco Soares[4] and Balthasar Weitzel[6]

[1]*Fraunhofer-Institute for Industrial Engineering IAO, Stuttgart, Germany*
[2]*ifeu – Institut für Energie- und Umweltforschung Heidelberg, Heidelberg, Germany*
[3]*University of Stuttgart, Institute of Human Factors and Technology Management IAT, Stuttgart, Germany*
[4]*Fraunhofer IAO, Stuttgart, Germany*
[5]*Municipality of Solingen, Solingen, Germany*
[6]*Fraunhofer Institute for Experimental Software Engineering IESE, Kaiserslautern, Germany*

ABSTRACT

Data Analytics projects often suffer from low data availability, inconsistent data formats and data models. In mobility data projects this especially means there are inconsistent attributes for specifying time and space dimensions. Also, non-data management and data analytics experts have so far had little opportunity to obtain an overview of data products, since data repositories and catalogs usually do not present the data in a self-describing manner. Data mostly exists in organizational silos as well as inadequate quality. In organizational departments data literacy is often focused on the own work area and does not include an overreaching view on data management. Arguing from a holistic view including data analytics as well as data management knowledge, here a "data analytics – data excellence gap" exists. This paper describes and motivates a smart city and mobility data ecosystem data catalog as a core component to bridge this gap.

KEYWORDS

Data Catalogue, Data Product, Data Analytics, Mobility, Real World Objects, Smart City

1. INTRODUCTION

Data is already used in various forms for everyday work in municipalities, research organizations and enterprises. In municipal as well as research and enterprise contexts the effort to collect and maintain data often outweighs the perceived added value. In general, in data analytics projects "data cleaning and preparation [even] takes approximately 80% of the total data engineering effort" (Zhang, S. et al 2003). From a data management point of view, there is a consensus in the scientific community that data quality (DQ) as "fit for purpose" (Wang and Strong 1996) for subject areas (DAMA 2017, p. 75) across departmental silos can only be achieved with data cleansing, master data management, and data quality management within organizations (DAMA 2017, Lis and Otto 2021, Sautter et al 2018). To deliver a solid and broad basis for decision, high quality data must be made available to others. In the discipline of research data management, and increasingly in other data management fields as well, it is becoming established that it should be kept FAIR, i.e., findable, accessible, interoperable, and reusable (European Commission 2021, Wilkinson, M. D. et al. 2016).

Coming from a citizen perspective, there are many digital services like social media platforms and online shops offering an excellent user experience. However, for beginner data scientists or civil servants, only a few possibilities exist to assess existing data products with a similar user experience.

Arguing from a holistic view, including data analytics as well as data management knowledge, a “data analytics – data excellence gap” exists: The enormous potential of FAIR as well as high quality data is far from being exhausted in data analytics projects, especially in mobility and geospatial analytics topics. However, as people in charge of data analysis tasks often are focused on their own domain of expertise only, they mostly have insufficient knowledge about broad and overarching data management. Thus, e.g., planning processes are often subject to the subjective experience of departments and resulting actions can be based on inadequate data. As a result, for instance in smart cities, the accurate estimation of future space requirements and necessary infrastructure becomes almost impossible and development stalls.

The MobiDataSol concept aims for designing a multiple organizations data trust model to make high quality data available using ecosystem data governance (Sautter, J. et al 2022). This paper serves as a subsequent step towards mobility data products provided by a Trusted Data Intermediary ecosystem that connects multiple organizations of municipal administration, research, and industry. The main paradigm, from a subject-matter data analytics point of view, therefore is: “Build and assess many small intermediate data products, document them, and hold them available also for others”. A data catalog, motivated and proposed in this paper, serves as a basic tool towards those many “intermediate data products” as well as an “intelligent data product” for Greenhouse gas accounting use cases fulfilling the project’s data analytics purpose. This paper first summarizes related work on data management and data analytics concepts, second describes the research method followed by section explaining the smart city mobility case study. A main section elaborates requirements for a data catalog in five subsections, before a short section summarizes the envisioned tool from a feature point of view.

2. RELATED WORK

Data Management Data products are the result of a value-added handling of data for data consumers (Wang and Strong 1996). From a technical and research point of view, two concepts for similar purposes aiming for storing and holding (and integrating) data can be distinguished: data lakes and data platforms (Schieferdecker, I. et al 2018) as well as data catalogues (Korte, T. et al 2019) and (meta-) data repositories (European Commission 2021). The former makes data directly accessible; the latter only provides metadata and links to the actual data sets hosted somewhere else. The boundaries between data lakes and metadata repositories can be blurred, depending on what is still interpreted as part of the actual data set and what is interpreted as metadata.

A data catalogue is understood as an integrated technical platform for data curation that aligns data supply and demand. It provides data inventory (data supply) and data discovery (data demand) functions. Further optional functions are data governance, data assessment and data analysis. Thus, a data catalogue is a critical tool for organizations seeking to optimize their data management practices.

When looking for data repositories and platforms with good user experience, the platform kaggle.com offers qualified metadata, a bronze, silver gold rating as well as user ratings for their open dataset allowing JSON, CSV SQLite and more datatypes[1]. Relevant platforms for mobility data in Germany are Mobilithek and research repositories situated at libraries like FID Move[2].

Metadata are information on data which can be subdivided in three subcategories: (1) descriptive metadata for identification purposes, (2) structural metadata on structure, attributes, and versioning as well as (3) administrative metadata for methodological and technical aspects related to data creation as well as access rights (Zeng 2004). An important initiative to enable better searchable and thus reusable public sector data is the DCAT-AP specification (European Commission 2022). It offers a unified data catalogue vocabulary that is the basis for a fine-grained search across national data portals within the EU. The actual search capability is realized by a hierarchical metadata harvesting mechanism, so that the European Data Portal[3] can be used to retrieve references to all Open Data published by public authorities of the EU.

[1] www.kaggle.com/datasets
[2] www.mobilithek.info, www.fid-move.de
[3] https://data.europa.eu

The processing of personal data requires compliance with data protection and privacy laws like the General Data Protection Regulation (GDPR) (EC 2016).

Data Analytics The CRoss Industry Standard Process for Data Mining (CRISP-DM) process model provides a step-by-step approach to data mining and data analytics processes, divided into the phases of business understanding, data understanding, data preparation, modelling, evaluation, and deployment (Wirth and Hipp 2000).

A major challenge in (big) data analytics is heterogeneity, "however, machine analysis algorithms expect homogeneous data, and are poor at understanding nuances. In consequence, data must be carefully structured as a first step in (or prior to) data analysis" (Jagadish et al 2014). Generating the right metadate to describe the recorded data is fundamental. E.g., in an experiment it is fundamental to offer details on the experiment conditions and procedures conducted. A special challenge in mobility data is representing the measures regarding space and "to acquire, store, analyze and visualize these datasets" (Demiryurek et al 2009). The data model shows the important key recorded variables such as sensor id, measure, space (i.e., coordinates) and time (timestamp), which allows further data aggregation such as hourly values, weekly values (cf. ibid.). To address these topics, among others in the domain of traffic information, the European Committee for Standardization developed DATEX II standard[4] for exchanging traffic information between traffic information centers, traffic operators and others, like media.

Regarding spatial information, the European union developed a nomenclature (NUT) to classify European regions into harmonized zones which can be easily referenced across datasets (Eurostat 2021). Other challenges mentioned are data inconsistency and incompleteness, scale (specific to big data), timeliness, and Privacy and Ownership (cf. ibid; Wang and Strong 1996).

Furthermore, a work conducted by Ardagna et al (2016) indicates Accuracy as a data quality issue: "To be aware of the accuracy level of a dataset before processing it. This problem may be even stronger with nonstationary data series.". It also refers to Usability and states that "On one side, data usability may increase with the availability of a more compact description of the dataset". Finally, it mentions challenges regarding Trust and Provenance "Accuracy and usefulness of the results in fact depend on the quality and origin of data, which in turn contribute to increase the trust of the final users in the process producing such results" (Ardagna et al 2016).

According to (Munappy et al 2020) data pipelines support solving data accessibility challenges by enabling data controllers to grant teams permissions to access data, originated from heterogeneous data sources at any point of the pipeline, which means at any level of (pre-)processing, granularity or enrichment as needed or possibly accessible (cf. data catalogue features). This eliminates the need for repeated collection and duplicated storage of the same data by multiple teams so sharing among them is facilitated. Data pipelines can be applied for data acquisition, data management, data quality monitoring, and machine learning applications (ibid.).

3. METHOD

First, we used own experience and consolidated related work on data management as well as data analytics requirements and solutions. Second, we identified data sources and gained subject-matter insights in semi-structured interviews with seven civil servants and executives as well as representatives from data-related enterprises in the city ecosystem of Solingen. The interviewees are experienced in the field of data analytics and/or manage relevant data regarding CHG. Roles of the civil executives and servants are Office of Geospatial Information, Statistical Offices, Mobility Planning, and Technical Operations. Beyond the identified locally available data sources, external information regarding transport data has been drawn together. Based on analytical requirements, a metadata scheme evolved (cf. 5.5). The data pipelining process that aims for elaborating a GHG accounting data product evaluates and determines the usability of the data catalog from a practitioner's point of view. Finally, both the final intelligent greenhouse gas data product as well as the data products may be (re-)used by various stakeholders in various use cases (cf. section 4).

[4] https://www.datex2.eu

Thirdly, we designed a data pipeline, which will later be technically implemented, in the form of a data flow diagram to define the necessary merging and calculation steps in the correct order. As a result of the processing steps leading to the final data product, a road-specific greenhouse gas accounting with the highest possible update frequency, intermediate data products are created, such as a traffic flow model, which can be relevant for further applications, for example in urban planning. Due to the relevance of these intermediate products for the quality of the final data product, additional requirements arise for the metadata of the data catalog. Fourth, we transferred the made conclusions in the case study to a general conclusion regarding data catalogs for data analytics.

4. CASE STUDY GREENHOUSE GAS DATA PRODUCT

In the information age, it is becoming increasingly challenging for municipalities to fulfil their tasks in an appropriate and future-oriented manner. Every municipal department has its own applications – be it the residents' registration office, the foreigners' registration office, the environmental department, the civil engineering office or the statistics department. These departments sometimes collect their data themselves or commission external service providers as e.g., research institutes, to create reports. However, it almost always leads to minimalistic data management and data use for a single purpose. In municipalities, Geoinformation Services typically are an exception as they offer services to other departments. Furthermore, initiatives driven by Geoinformation departments may exist that strive for holding and managing data for multiple-purposes for cross-department needs. Often, however, these fail due to a shortage of personnel. GIS systems and geodata departments fulfill the function of collecting data from different departments (e.g., using linked open data portals), because there is a legal obligation to publish and version some types of data. Geospatial data, like maps for example, are accessible and easy to understand. But by no means this represents the whole picture of what is realistically possible or already implemented by municipalities.

Previous approaches for traffic and Greenhouse gas emission estimations on a regional level (municipality, county, or administrative districts) are often based on a rather rough or outdated estimation of the traffic volume and sources on site, which also lead to an unprecise and not up to date estimation of traffic-induced emissions. Greenhouse gas accounting, deriving measures and monitoring could be carried out for different relation types like source-destination, domestic or commuter traffic more accurately and specific. During the development of a basic high-quality database, that provides all the available data of the and regarding the local transport system could be used by many actors involved in environmental monitoring and transport planning.

A Data Catalogue is the core component to bridge between the data-based ways of operating a smart city and everyday challenges of municipalities – or to put it differently: to bridge the Data Analytics – Data Excellence Gap in the public context.

5. REQUIREMENTS FOR A DATA CATALOG

In Data Governance projects aiming to consolidate an organization’s data excellence the introduction of a data catalog (DC) to improve the transparency of data types, qualities and quantities is a typical early step. From an (expert- and non-expert) user point of view, both is needed: Ensuring data quality as well as offering subject matter quality e.g., in mobility. Also, the DC should comply with high user experience standards.
From a data analytics point of view, the systematic documentation and description of intermediate steps and even making data products available to other (external) stakeholder often comes up too short. Further, main data management paradigms, such as “Get usage right also for second usage before you acquire or assess data” is often not considered, especially in municipal and research contexts. Regarding personal data, the minimization of storage is an important issue to consider. In general, the use of anonymous data is preferable since in this case the data protection requirements do not apply.

In the following specification, we distinguish the data product, the physical object of the real world (e.g. a vehicle) and the data object (e.g. table row/line in the data model/stream of the data product). Requirements for a data catalog described in the following reach beyond classical metadata specifications in the eGovernment, research data and geospatial field[5] and data catalog or data repository features (cf. section 2).

5.1 Bibliographic and Identifying Metadata for the Data Product

Regarding metadata (data describing data), we distinguish between bibliographic and identifying metadata and content descriptive metadata. Data for the purpose of identification (e.g., title, abstract/summary, author and keywords) and methodological texts that are recorded at the level of the data set. Information on the origin, collection period and nature of the data.

Connecting applications to the data catalog can further enhance its value in two keyways. Firstly, metadata contained within the data catalog can be leveraged in business intelligence or advanced analytics applications, enhancing the accuracy and completeness of analytical results. Secondly, by integrating external systems or data sources, data catalogs can facilitate user collaboration and streamline data governance workflows. Bibliographic and identifying metadata requirements are:

- Persistent Identifier
- Curator/Reviewer
- Data Provider (organization/geodata service/integration level/data warehouse)
- Data Consumer (and known Quality requirements)
- Data Privacy Level
- Data Type/Format/Data Model (i.e. DATEX II)
- Availability/Licensing
- (Restricted/Unrestricted)
- Certificate terms/Maturities of certificates
- Dependencies to other data products/sources

Further requirements are given by the metadata standards DublinCore, DataCite (research data), INSPIRE, OCG (geodata) and DCAT-AP (eGovernment/public data).

5.2 Object Reference

Spatial reference: Ideal, for the case of the EU and UK, usage of NUTS (Eurostat 2021) nomenclature up to Local Administrative Units (LAUs are the building blocks of the NUTS which comprise municipalities and communes of the European Union) and further sub census division levels. Time reference: to which date/ period does the data refer to. Here, things become complex because observations should be allocated to a standardized datetime format, even though they might refer to a certain period. For this, ideally there is a timestamp which indicates the beginning of the period and a further data feature which indicates the duration of that observation e.g., if we are talking about a monthly value then we need to make some standardized way to set it; one way is to set the value to the first day of the corresponding month (if year, first day of the corresponding year) and add a variable "period" that indicates the duration.

Unit reference: Another important reference is the observed entity. In the smart city context, traffic can be viewed from different perspectives. On vehicle level or on street and district level, do we view traffic as a whole or do we distinguish between cars, buses, trucks or bikes. This point is related to both spatial reference and aggregation level. Nevertheless, it deserves mention in this section.

In these cases, it is also important to indicate if and how the observed value was aggregated (e.g., average of multiple observations, sum of multiple observations, etc.). Therefore, in general, units should be properly indicated and contextualized: we should follow the SI system (when available) and in case of aggregated data this information should follow (e.g. kg/vehicle). The system of units used, and the sizes must always be indicated and if necessary a conversion to SI system should be offered. When data is aggregated, the aggregation method should also be clearly described (in the variable metadata?)

[5] https://datacite.org/, https://www.dcat-ap.de/, https://www.dublincore.org/ , https://inspire.ec.europa.eu/

5.3 Object Measurement

Another important information is the source of the data: the data provision system (e.g., automated traffic acquisition system) and the data provider (e.g., municipal traffic office), as well as the available technical documentation for data acquisition and processing, should be fully available in their most current versions. It should be clear what devices/tools were used for the data collection (do they follow a standard as e.g., ISO) or at least their capabilities and restrictions. Techniques/methods for calculating the final value (e.g., if the specified value is the altitude but this is measured with a pressure sensor) must also be described, assumptions made must be mentioned. In addition, it is important to know in which intervals the data were collected. Is it a one-time survey or are the measurements performed in a regular cycle (daily, monthly, annually) and will they be continued to be recorded in the foreseen future? In order to be able to assess the accuracy of an analysis, the uncertainties and reliability (e.g., failure rates) are crucial..

5.4 Data Product Representativeness

Finally, it is important to understand the representativeness of the collected data: What was the sample of the data and what does it represent? (i.e., on surveys, what was the sample? Does it represent the spatial population composition?) Representativeness might not be mandatory but information whether it is, should be indicated particularly if there was a special scope of the collection as e.g., public transport users.

5.5 Data Product Content Describing Metadata and Status

Under content-describing metadata, we understand characteristics, structure and versioning of structured data (e.g., columns of a table), which are either described at file/data stream level (e.g., file type, raw date vs. processed data) or contained in files. In addition, information is required about the status of the data within the data pipeline (reference to data pipeline system cf. section 2). This means that, on the one hand, status messages must be forwarded from the processing chain that indicate errors. On the other hand, there are cases where an update of an intermediate product is not mandatory for the creation of a valid final product.

- Data product name
- Content summary and object description
- Data Type/Format/Data-Model (i.e., DATEX II)
- Geographical Scope (regional/national/global)
- Reference unit Scope (road/car/distance)
- Scope Means of transport (car/semi/bike)
- Boundary User group affiliation
- Data collection period
- Update frequency
- Foreseen availability (one-time vs. continuous survey or collection)
- Precision or uncertainties if available
- Data pipeline acquisition method (Push- vs. Pull-API)
- Aggregation of units/timeframes/spatial
- Suitability for analytical questions
- Analytical/Interpretation Restrictions
- Precise Description of key metrics and corresponding units
- Static or dynamic data (list of parking spaces vs. near real-time capacity utilization of parking spaces)
- Status of upstream data (products)
- Status of maintenance (“error has been recognized and will be fixed”)

6. THE DATA CATALOGUE BUILDING BLOCK

Summing up the requirements, key features of an envisioned data catalog solution can be summarized as follows:

1. Data cataloging for metadata and corresponding full data of intelligent data products (that may refer to other data products)
2. Integration with data analytics processes (data pipelining) and support for “intermediate” data products

3. Manage the reference to the real-world objects standing behind data catalogue items (reference, measurement and representativeness)
4. Inherent features for cross-organizational access right management and data governance allowing compliance and data quality.
5. User Experience for public access by citizens (open data platform) as well as internal users
6. Fostering global identification (also for internal data products) and global findability

Beyond this basic concept for a data catalog, the following other enhancing aspects are relevant to tackle a sociotechnical solution for (ecosystem) data governance: (1) a data governance role model/curator and data steward per data product (who cares if the (dynamic) data product is broken?), (2) a data pipeline (tool/process) used by data analyst (validity/quality etc.), (3) interoperability for data products that allows an automatic creation of data pipelines based on bibliographic metadata (4) an automatic update of metadata in case of measurement parameters/assumptions/boundaries/tolerances change (5) An easy-to use automatic or half-automatic integration with public data platforms as well as research data repositories[6]. (6) a data space ecosystem connector as well as (7) an integration with municipality, enterprise and research institute structures and processes.

7. CONCLUSION AND OUTLOOK

This paper focused on deriving requirements for a data catalog based on expert's insights as well as own case study experiences. As a contribution to the state of the art, we define four main data requirements: Object reference (where does the data refers to?), object measurement (how is/was the measured quantity collected?), data product representativeness (what does the data under which assumptions represent?) and data product status how valid and of which operational value is the data?).

In data analysis, especially in the field of mobility, environmental and spatial data analysis, aspects of validity, accuracy, representativeness as well as the coverage quality of the spatial and temporal value progression are guiding questions to be answered in the data preparation phase. Coming from a data product-oriented approach of data analytics especially in the field of mobility and smart city data, using such a data catalog with an inherent management of metadata, and thus the knowledge about data, could significantly reduce the efforts for data analytics projects. To foster the adoption of the data catalog in public, private and research contexts, a mapping to respective vocabularies is required. Next steps should tackle the question of how these data can be put into value.

ACKNOWLEDGEMENT

We kindly thank the whole MobiDataSol consortium, especially Dominik Lis and Nicolas Ortiz, who had an essential contribution to the work of the paper. The MobiDataSol project, funded by the German Federal Ministry of Education and Research (funding code 16DTM103B) and supervised by VDI/VDE, aims at elaborating a concept for ecosystem data governance for mobility data products in the context of smart cities. Funded by the European Union – NextGenerationEU.

REFERENCES

Ardagna, C. A. ; Ceravolo, P., and Damiani, E.: "Big data analytics as-a-service: Issues and challenges," 2016 *IEEE International Conference on Big Data (Big Data)*, Washington, DC, USA, 2016, pp. 3638-3644, http://doi.org/10.1109/BigData.2016.7841029.

[6] E.g. http://www.mobilithek.info, http://www.fid-move.de etc.

DAMA (Ed.), 2017. D*AMA-DMBOK: Data management body of knowledge (second edition ed.)*. Technics Publications, Basking Ridge, New Jersey.

Demiryurek, Ugur & Banaei-Kashani, Farnoush & Shahabi, Cyrus, 2009. "TransDec: A Data-Driven Framework for Decision-Making in Transportation Systems," 50th Annual Transportation Research Forum, Portland, Oregon, March 16-18, 2009 207726, Transportation Research Forum.

Dobrokhotova, Ekaterina; Engelbach, Wolf; Sautter, Johannes (2015): *Marktstudie 2015 Multidomänen-Stammdatenmanagementsysteme.* Stuttgart: Fraunhofer-Institut für Arbeitswirtschaft und Organisation (IAO), https://doi.org/10.24406/publica-fhg-297449.

Eurostat, G., 2021. *Statistical Regions in the European Union and Partner Countries. NUTS and Statistical Regions* 2021. Eurostat, p.188.

European Comission 2016. General data protection regulation (GDPR): Regulation (EU) 2016/679 of the European Parliament and of the Council of 27th April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC. http://eur-lex.europa.eu/legalcontent/EN/TXT/?uri=CELEX:32016R0679.

European Commission, 2021. *Recommendations on FAIR Metrics for EOSC.* Publications Office of the European Union, Luxembourg. https://doi.org/10.2777/70791

European Commission, 2022. *DCAT Application Profile for data portals in Europe*. Publications Office of the European Union, Luxembourg. https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe/release/211

Franklin, M. et al, 2005. From databases to dataspaces. *ACM SIGMOD Record* 34, 4 (2005), 27–33. https://doi.org/10.1145/1107499.1107502

Jagadish, H. V.; Gehrke, Johannes; Labrinidis, Alexandros; Papakonstantinou, Yannis; Patel, Jignesh M.; Ramakrishnan, Raghu; Shahabi, Cyrus (2014): Big data and its technical challenges. *In: Commun. ACM 57 (7)*, S. 86–94. https://doi.org/10.1145/2611567.

Korte, T. et al, 2019. *Data Catalogs - Integrated Platforms for Matching Data Supply and Demand: Reference Model and Market Analysis (Version 1.0)*. Fraunhofer Verlag, Stuttgart.

Kühling, J. 2021. Der datenschutzrechtliche Rahmen für Datentreuhänder. *Zeitschrift für Digitalisierung und Recht* 1 (2021*)*, 1–27.

Sautter, Johannes; Litauer, Rebecca; Fischer, Rudolf; Klages, Tina; Wuchner, Andrea; Müller, Elena et al. (2018): Beyond Data Quality: Data Excellence Challenges from an Enterprise, Research and City Perspective. *In: Proceedings of the 7th International Conference on Data Science, Technology and Applications.*, S. 245–252, DOI: 10.5220/0006912902450252, Available online at https://publica.fraunhofer.de/handle/publica/401256.

Sautter, Johannes; Lis, Dominik; Kräck, Jan; Helsper, Andreas; Erlenhardt, Kai; Schnieders, Fabienne; Lambrecht, Udo (2022): Mobility Data Products for Smart City Ecosystems – a Greenhouse Gas Balancing Case Study. In Yingcai Xiao (Ed.): International Conference on Connected Smart Cities (CSC2022), Lisbon, Portugal, 19-22 July 2022. Red Hook, NY: Curran Associates Inc. Available online at https://publica.fraunhofer.de/handle/publica/425148.

Schieferdecker, I. et al, 2018. *Urbane Datenräume - Möglichkeiten von Datenaustausch und Zusammenarbeit im urbanen Raum.(Urban data rooms – possibilities of data exchange and cooperation in the urban environment).* Fraunhofer FOKUS, Berlin. http://publica.fraunhofer.de/documents/N-500021.html

Wang, Richard Y. and Strong, Diane M. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems* 12, 4 (1996), 5–33.

Weber K. and Klingenberg C. (Eds.). 2021. *Data Governance: Der Leitfaden für die Praxis.* Carl Hanser Verlag GmbH & Co. KG, München. https://doi.org/10.3139/9783446466746

Wilkinson, M. D. et al, 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018 doi: 10.1038/sdata.2016.18.

Wirth, Rüdiger; Hipp, Jochen (2000): CRISP-DM: Towards a standard process model for data mining. *In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining.* Citeseer, S. 29–39. Online available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.198.5133&rep=rep1&type=pdf.

Lis, D. and Otto, B. 2021. Towards a Taxonomy of Ecosystem Data Governance. *In Proceedings of the 54th Hawaii International Conference on System Sciences*, Tung Bui (Ed.). , https://doi.org/10.24251/HICSS.2021.733.

Zeng, M. (2004). *Metadata types & functions*. http://marciazeng.slis.kent.edu/metadatabasics/types.htm.

Zhang, Shichao; Zhang, Chengqi; Yang, Qiang (2003): Data preparation for data mining. In: Applied Artificial Intelligence 17 (5-6), S. 375–381. https://doi.org/10.1080/713827180.