OPAL
OPEN DATA PORTAL

Deliverable D3.1
Spezifikation von Qualitätskriterien

Autoren: Adrian Wilke, Ivan Ermilov, Matthias Wauer

Veröffentlichung	Vertraulich
Fälligkeitsdatum	30.06.2018
Fertigstellung	30.06.2018
Arbeitspaket	AP3
Typ	Bericht
Status	Final
Version	1.0

Kurzfassung:

In der wissenschaftlichen Literatur wurden bereits unterschiedliche Aspekte zu offenen Daten und Metadaten untersucht. Dieser Bericht beschreibt die Ergebnisse eines hybriden Ansatzes hinsichtlich eines Systems zur automatischen Generierung von Qualitätsmerkmalen von Metadaten zu offenen Datensätzen im Bereich von Mobilitätsdaten. Der hybride Ansatz setzt sich aus einer Literaturrecherche und der Untersuchung von Metadatensätzen aus derzeit verfügbaren Portalen zusammen. Das Ergebnis ist ein Katalog, bestehend aus 13 Qualitätsdimensionen und 48 zugehörigen Qualitätskriterien und -metriken zur konkreten Implementierung.

Schlagworte:

Metadaten, Qualität, Mobilitätsdaten, Metriken, Offene Daten

mFUND
Das Startkapital für die Mobilität 4.0

Bundesministerium
für Verkehr und
digitale Infrastruktur

Inhalt

1 Introduction

Datasets in Open Data portals often are only valuable for applications if their metadata matches certain criteria. Hence, the quality of this metadata is a primary concern for the usability of such datasets. In this deliverable, we therefore specify quality criteria by which the datasets and their respective portals can be judged.

1.1 Motivation

Our hypothesis is that metadata of datasets should conform to certain quality criteria in order to be useful. An example for such a criteria is the availability of licence information. Users will only consider a dataset for their applications under certain conditions, e.g., that the licence allows commercial use. This criterium can be graded further. For instance, the criteria could span the range of the availability of any licence information, such as a literal value ("Creative Commons Share-alike"), to having a machine-interpretable reference to licence details (e.g., modeled using an ontology[1]). Furthermore, the criteria can be represented by metrics, which portrait certain aspects of metadata quantitatively. Thus, they allow a comparison of datasets or, in aggregated form, portals.

1.2 Methodology

We approach the problem of metadata quality criteria and metrics (MQCM) definition from two perspectives. The first one is the analysis of the state-of-the-art research in the metadata quality research. We call the first perspective a top-down one as it starts from the conceptual analysis of the problem. We have analyzed the relevant state-of-the-art papers on data quality as well as related projects (e.g. ADEQUATe). The second perspective is inferencing MQCM from the data analysis we performed on data from mCLOUD and European Data Portal. The second perspective is called bottom-up as we start from the data itself. Finally, the results are aggregated in a table, which will serve a requirements specification for Civet, a quality analysis framework developed in OPAL.

[1] Oleksandra Panasiuk, Simon Steyskal, Giray Havur, Anna Fensel and Sabrina Kirrane: Modeling and Reasoning over Data Licenses, ESWC 2018.
https://2018.eswc-conferences.org/files/posters-demos/paper_298.pdf

2 Related work

Besides algorithms, the underlying data has become a topic of interest also for non-specialists in the last decade. In this section, we provide a brief overview of research in the fields open data, metadata, linked data and the intersection of the topics as well as related quality criteria, which are relevant for the OPAL approach. A detailed description of the quality criteria in the following are listed in the appendix of this document.

Open data

Regarding open and mobile data, which is the focus of the OPAL project, there are the *8 Principles of Open Government Data*, published in 2007. These principles represent basic properties of providing data and therefore a set of quality criteria. In short, the resulting principles are: complete, primary, timely, accessible, machine processable, non-discriminatory, non-proprietary, license-free. The principles can be accessed online[2]. (See appendix A.G.)

Tim Berners-Lee, often referenced as the inventor of the World Wide Web, presented the idea of *5-star Linked Data*[3]. This is a list of criteria for data formats, build upon each other. The criteria are: The choice of an open licence for publishing data, using a machine-readable structured format (e.g. no PDF), additionally use a non-proprietary format (e.g. CSV), choose open standards from W3C (like RDF), and create linked data (to provide the option to aggregate different repositories). In OPAL, metadata of different sources has to be transformed to meet all of the requirements. (See appendix A.L.)

Zaveri et al.[4] performed a survey and filtered 21 papers related to data quality, dimensions and metrics for 10 years up to 2012. They worked out 23 dimensions for Linked Open Data quality and presented each dimension by giving a definition, metrics, descriptions, and an example. In short, the aggregated dimensions are: accessibility (availability, licensing, interlinking, security, performance), intrinsic (accuracy, consistency, conciseness), trust (reputation, believability, verifiability, objectivity), dynamicity (currency, volatility, timeliness), contextual (completeness, amount-of-data, relevancy), representational (representational-conciseness, representational-consistency, understandability, interpretability, versatility). The presented extensive catalogue of quality dimensions and metrics was worked out to evaluate open data, but the approach is not specialized for the particular case of open metadata. (See appendix A.Z.)

Open metadata

The *FAIR*[5] data principles have been worked out to support the discovery, evaluation, and reuse of data by humans and machines. To meet the four FAIR principles, data has to be findable, accessible, interoperable, and reusable. The principles comprise the use of protocols, identifiers, licences, and standards. Altogether, FAIR specifies a set of criteria for both, data and metadata,

[2] Aaron Swartz et al. (2007). 8 Principles of Open Government Data. http://web.archive.org/web/20071214024243/https://public.resource.org/8_principles.html

[3] Berners-Lee, T. (2010). Linked Data - Design Issues. http://web.archive.org/web/20101202183255/https://www.w3.org/DesignIssues/LinkedData.html

[4] Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for Linked Data: A Survey: A systematic literature review and conceptual framework. Semantic Web, 7(1)

[5] Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3. https://doi.org/10.1038/sdata.2016.18

but lacks in providing concrete metrics or specifications on the implementation of the criteria. (See appendix A.W.)

For the special case of open government metadata, Reiche et al.[6] defined the following 7 metadata quality metrics: completeness, weighted completeness, accuracy, richness of information, readability, availability, and misspelling. For each of the metrics, a formula is provided, which can be used as a basis for implementation. The authors state, that the quantification and the algorithmic approach is too limited to discover all subtleties that result in quality flaws. The work is strongly related to the development of the German Open Data Portal (govdata.de). (See appendix A.R.)

Based on the quality criteria of Reiche et al., Umbrich et al.[7] created a catalogue of 6 quality dimensions (retrievability, usage, completeness, accuracy, openness, and contacability) and metrics to automatically calculate them. Based on these definition, they conducted a comparison of quality of 82 CKAN portals. Additionally, the results of the implementation can be accessed on Open Data Portal Watch (http://data.wu.ac.at/portalwatch/). (See appendix A.U.)

The work was continued in the ADEQUATe project (http://adequate.at/). Neumaier et al.[8] refined the metrics into 5 categories with 18 quality dimensions (existence, conformance, retrievability, accuracy, and open data) and respective metrics. Additionally, they integrated three portal software implementations (CKAN, Socrata, OpenDataSoft) into the Data Catalog Vocabulary (DCAT) and conducted a quality evaluation of 261 open data portals. (See appendix A.N.)

We use the presented works of research for the specification of metadata quality criteria and metrics in the following sections.

[6] Reiche, K. J., Höfig, E., & Schieferdecker, I. (2014). Assessment and Visualization of Metadata Quality for Open Government Data. In Conference for E-Democracy and Open Governement (pp. 335--346). Krems.

[7] Umbrich, J., Neumaier, S., & Polleres, A. (2015). Quality Assessment and Evolution of Open Data Portals (pp. 404–411). IEEE. https://doi.org/10.1109/FiCloud.2015.82

[8] Neumaier, S., Umbrich, J., & Polleres, A. (2016). Automated Quality Assessment of Metadata across Open Data Portals. Journal of Data and Information Quality, 8(1), 1–29. https://doi.org/10.1145/2964909

3 Metadata Quality Criteria and Metrics

3.1 Definitions of quality

Data quality is commonly conceived as fitness for use for a certain application or use case. The scope of the OPAL project is German Open Data not specific to any domain, but usually related to mobility data. For such open data, we define the data quality empirically via a set of quality criteria.

3.2 Quality criteria

In this section we define quality criteria, which show when the open data is fit for use by the stakeholders (e.g. citizens, companies, research institutions). We conducted a literature review based on the related work articles (see Sec. 2). Therefore, we examined the quality criteria formerly used in scientific works (see tables in the appendix) and clustered overlapping criteria descriptions. This resulted in a list of 67 selected and pre-aggregated criteria.

The list was reviewed for the purposes of the OPAL project. We selected and refined those criteria, which fit in the domain of open mobility metadata. Criteria, which can only applied by accessing data records as well as criteria, which cannot be quantified, were not included. This produced 12 quality dimensions with 43 consolidated quality criteria. The resulting matrix of the literature review (see Appendix A.A) will be extended with results of the data-driven investigation in Section 4.

3.3 Quality metrics

An automatic processing of metadata records and a related generation of quality criteria values can be implemented by a quantification of the quality criteria. For the consolidated criteria list in the previous section, we added descriptions of possible metrics in relation to the previous scientific works for the upcoming implementation of the OPAL system. The metrics were added to the table in Appendix A.A.

4 Data-driven investigation

In this section we describe the datasets used for the extension of the quality criteria of the previous section. We provide basic statistics for the datasets as well as quality criteria as defined in Section 3. The additional specific criteria based on the focussed OPAL resource portals will be added to the final catalogue in Section 5.

4.1 European Data Portal

The crawl of the European Data Portal was done in February 2018 and contains 804,982 datasets[9]. In the OPAL project we address the problems of German government Open Data, thus we filter out German datasets from the portal, which results in 219,768 German datasets[10].
In the following we list the criteria inferred from the data analysis performed on European Data Portal. It is important to note that OpenDataMonitor quality criteria are covered by the criteria below.

No.	Criterion	Description
1	Missing Values	Amount of missing data, which should be present in the schema defined by a data portal
2	Non-conforming Metadata	Metadata provided is non-conforming any of the existing standards or is stored in a field-extension (extra field), which is not described by any existing specifications
3	Non-existing Ontologies	Metadata is using non-existing or non-dereferenceable ontologies, which makes it hard to understand the meaning behind resource URIs
4	Contact Point Completeness	The contact point is stated clearly including an organization generated the metadata, maintainer's email address, telephone number and address.
5	Versioning Information	Metadata clearly indicates the time period for the data gathering, the affected geographical region (if applicable) as well as version of the dataset (if applicable)
6	License	Can I use this dataset for academic purposes? How easy it is to understand the license? Is the common licence used (e.g. CC-BY, CC0, or some very well-known governmental license?)
7	License Commercial	Can I use this dataset for commercial purposes?
8	Easy-to-understand Description	Analysis of German language to state difficulty level → e.g. this text can be understood by (graduates, school students, children) etc.

Table E: Data Quality criteria from the data analysis of EuropeanDataPortal.

4.2 mCLOUD

The size of the mCLOUD metadata corpus (856 datasets on June 28th 2018) is not appropriate for a meaningful statistical analysis as part of the data-driven investigation. Therefore, we apply the random sampling approach and use the sample of OPAL deliverable D1.2. Table M.D shows an overview of the used dataset samples.

[9] https://www.europeandataportal.eu/

[10] https://github.com/earthquakesan/odp-metadata-analysis/blob/master/ODPAnalysis.ipynb

No.	Title	Provider	Link
1	Jährliche Raster von Winterraps - Beginn der Blüte in Deutschland	Deutscher Wetterdienst (DWD)	mCLOUD
2	Monatliche Sonnenscheindauer	Deutscher Wetterdienst (DWD)	mCLOUD
3	Serviceeinrichtungen	DB Netz AG	mCLOUD
4	RadwegeGis Hamburg	Hamburg: Behörde für Wirtschaft, Verkehr und Innovation, Amt für Verkehr und Straßenwesen	mCLOUD
5	Grundwassergleichen Max 2008	Hamburg: Behörde für Wirtschaft, Verkehr und Innovation, Amt für Verkehr und Straßenwesen	mCLOUD
6	Bund: Farbrelief des Wasserlaufs	Informationstechnikzentrum Bund (ITZBund)	mCLOUD
7	VBB-Fahrplan 2013	VBB - Verkehrsverbund Berlin-Brandenburg GmbH	mCLOUD
8	Urbane Räume: Lufttemperatur und Luftfeuchte stündlich	Deutscher Wetterdienst (DWD)	mCLOUD
9	Digitale Bundeswasserstraßenkarte im Maßstab 1:1.000.000	Generaldirektion Wasserstraßen und Schifffahrt (GDWS)	mCLOUD
10	VBB-Fahrplandaten August 2017 bis Dezember 2017	VBB - Verkehrsverbund Berlin-Brandenburg GmbH	mCLOUD

Table M.D: Random sample of mCLOUD datasets based on OPAL D1.2

The mCLOUD metadata samples have been manually analyzed. Both, included quality criteria as well as missing quality criteria, have been collected and are listed in the following Table M.Q. The table additionally contains examples and solutions to quantify the criteria, if they can not obviously be conducted.

No.	Criterion	Description with example and description of metric, if necessary
1	Locality	Is geographical information available in the title or description? (ID of positive example: 1, ID of negative example: 3)
2	Extended description	Does the description provide additional information in comparison to the title? (Pos: 5, neg: 1) An automatic computation can be archived with a word comparison.
3	License	Is the name of the related license given? This opens up the possibility to investigate the rights of use. (Pos: all, neg: none)
4	License link	Is there a link to the license to identify it in an unique manner? (Pos: all, neg: none)
5	Openness	Are the given datasets open? (Single regulations like giving attribution to original dataset, licensing of derived data, commercial use can be refined) (Pos: all - but limited in single aspects, neg:none)
6	Multiple formats	Are links to multiple, content-specific formats given? For mobile data, e.g. Web Map Service (WMS) and Web Feature Service (WFS) could be provided. (Pos: 4, neg: 8)
7	Accessible	Are the given links accessible (or in other case not available online)? (Pos: all, neg:none)
8	Timeliness of dataset	Is the point of time of the last update of the dataset given? (Pos: all, neg:none)
9	Timeliness of dataset description	Is the point of time of the last update of the metadata entry given? (Pos: all, neg:none)
10	Categorization	Are categories or tags for classification given? (Pos: all, neg: none)

11	Category links	Are there links or URIs to identify and combine the categories? (Pos: none, neg: all)
12	Provider	Is the provider of the dataset listed? (Pos: all, neg: none)
13	Provider contact	Is there information to contact the provider? (This can be refined, links could be general or specific, a contact email address or telephone number are more direct ways to contact.) (Pos: all - with general link, neg: none)
14	Data formats	Is the format of the related dataset well-described? (For structured data formats, the format itself or rather the documentation of the data format provides satisfying information) (Pos: 4, neg: 1)

Table M.Q: Data Quality criteria from the data analysis of mCLOUD.

mCLOUD metadata description texts

mCLOUD metadata consists of structured (e.g. last time of update) and unstructured data (e.g. description texts). For refinements and extensions of single metadata records, unstructured data can be analyzed to extract entities and combine them with additional information. Therefore, the scope of description texts can be considered. For the extraction of entities out of description texts, the number of provided words as a quantitative measure can be important to implement quality heuristics for the upcoming augmentation of available data. Therefore, the mCLOUD portal was accessed on January 29th 2018 and 652 datasets have been extracted. An overview of the frequency of words in description texts is presented in Figure 1. There was no description text found for 5.71 percent (in total: 49) of the datasets. Descriptions between 1 and 10 words are provided 21.38 percent (in total: 183). Summarized, the majority of mCLOUD metadata entries provide sufficient data for an entity recognition to augment and refine data.

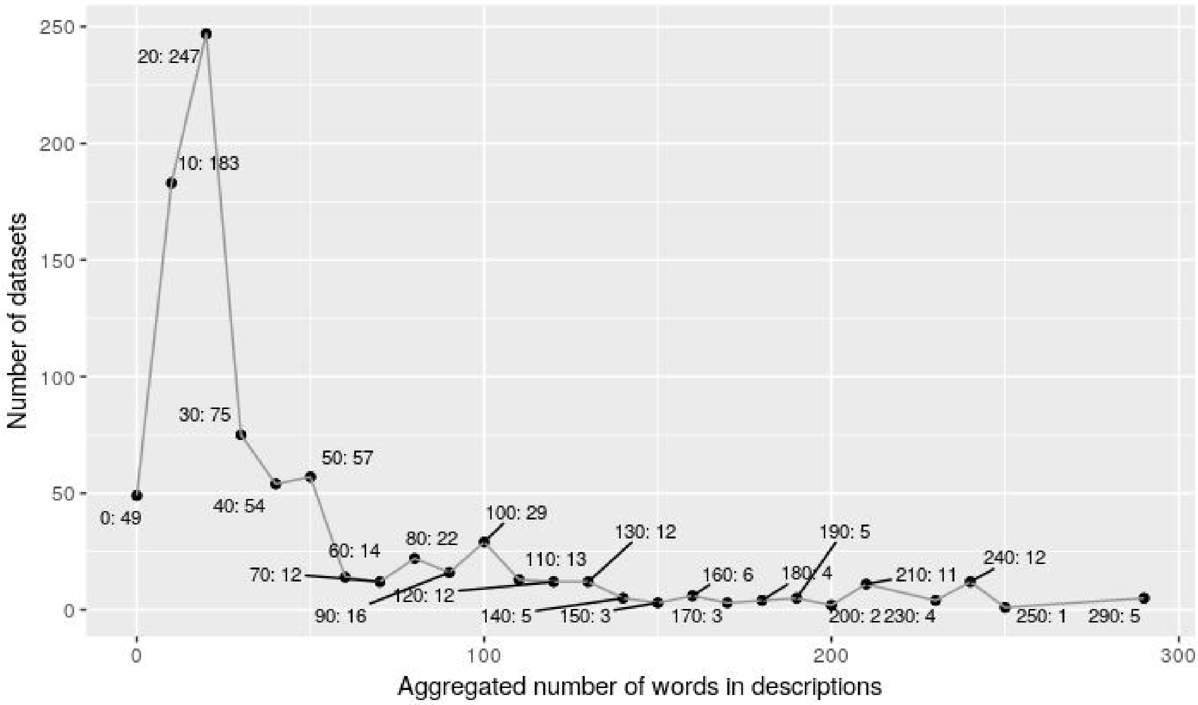

Figure 1: Frequency of numbers of words in mCLOUD descriptions

5 Specification of quality criteria

Based on the results of the literature review (top-down approach) and the data driven investigation (bottom-up approach) we present final quality criteria and metrics.

5.1 Generic and specific quality criteria

Previous works on quality regarding open data often focus on the respective research interest. In the literature, works on Open Data, Open Metadata, Open Government Data, Linked Data, and FAIR data were found. The research focus of the OPAL project is on an overlapping field. Metadata records regarding open datasets, related to mobility data, and often published by government offices form the data basis of the system to develop and implement. Based on the literature and real world metadata records, we identified the following quality dimensions:

- **Expressiveness**: The totality of metadata fields should express the content and context of the related data.
- **Temporal**: Metadata should be updated, if new versions of the related data are published.
- **Understandability**: A clear language and example applications support the understanding of the described metadata.
- **Rights**: The choice and the description of licenses fundamentally open or restrict the options for data usage.
- **Trust**: The provider of related data is an indicator for data quality.
- **Community**: Discussions, ratings, and referrals of data users can augment the context of datasets.
- **Versatility**: Metadata can be represented in different formats, accessed by several interfaces, and expressed in various natural languages.
- **Representation**: The structure of metadata should be formed in a way to support an open, human and machine friendly usage.
- **Interlinking**: Linked data provides options to describe data across structural boundaries.
- **Contactability**: Contact information opens up possibilities to handle errors inside data and to drive data advancements.
- **Access**: Metadata should be accessible in an open, direct manner.
- **Data**: Besides criteria on the metadata, some information inside metadata can be used as indicators for the related data records.

These dimensions are refined into concrete criteria and metrics in the next section.

5.2 OPAL Quality Criteria and Metrics for Open Metadata for Mobility Data

The 13 resulting quality dimensions and including 48 criteria and metrics, based on the works of this deliverable are presented in the following table:

No.	Criterion	Description	Metric
Expressiveness			
1	Extend	A metadata entry for a specific dataset should consist of an extensive, non-empty set of metadata fields.	Count the number of given metadata values and divide it by the total number of used metadata fields.
2	Weighted extend	For different fields of interest, specific metadata fields can become a related importance. This can be quantified by applying (user-based) weights for individual metadata fields.	Multiply given metadata fields with a weight, compute the sum of all weighted values, and divide it of the largest possible value of a sum.
3	Categorization	Especially categories and tags can be used to discover new datasets and vice versa a current dataset can be discovered from other sources.	This can be implemented by checking a true/false value, if a metadata record contains tags/categories or not. Additionally, the number of categories and tags can be used for quantification.
4	Description*	In real world metadata, description text may be very short or consist only of a copy of the title.	Check, if the description text is non-empty, does not overlap with the title, and extends N words.
Temporal			
5	Timeliness	Data users often are interested in up-to-date data.	Calculate the difference of the current time and the time of creation or the last update. Only one timestamp in the metadata is needed.
6	Update rate	Information about the frequency of updates.	Sum up the number of update events.
Understandability			
7	Readability	The readability of text can be estimated by sentence length and parts of speech.	Usage of a readability test, e.g. Flesch-Reading-Ease
8	Language errors	Correct language in metadata fields is can be an indicator for data quality, especially if parts of the metadata have been extracted from the original data.	Count the number of words written incorrectly. Additionally, heuristics for grammar errors can be applied.
9	Example applications	Example applications, which are based on data described by a metadata record can be software projects or articles.	Check, if there are links to examples.
Rights			
10	Machine readable license*	A machine readable license can be processed automatically to provide and aggregate information for users. For this, IDs like official abbreviations or URLs of official websites can be used.	Check, if there is an ID, URL, or structured information about the license.

11	Human readable license	In user interfaces, listed licenses can provide users with information about rights or at least provide the option to look up the related license permissions and restrictions.	Check, if the metadata contains a title of the license.
12	Known License	Properties of well-known licenses can be represented in different ways to users and are an added value.	Check, if the license is listed in a license database.
13	Open License	Open licenses typically extend the range of possible applications for work with the related data.	Check, if the license is described as open.
14	Permission for commercial use	Information extracted out of data is valuable and can be integrated in commercial databases, applications or just sold to reward the work with data.	Check, if the license is described as free for commercial use.
15	Permissions*	Some datasets are restricted to give attribution to the original author. In other cases, the data may be available for academic purposes. The awareness about these restrictions are important for re-publishing data.	Check, if the license meets properties like attribution, share-alike, or public-domain.
Trust			
16	Provider identity	Data can be provided from instances like government, universities, or arbitrary humans or machines. A known identity can be used for researching additional information	Check, if the official name, title, or website of the data provider is provided.
17	Trusted provider	The credibility and reliability of the metadata provider can be stored in a database and metadata records from the same provider can be rated.	Check, if the metadata provider is listed in a provider database with rating information.
18	Metadata authenticity	If several publishers or sources for one metadata record are found or if there are overlappings among data contributors for several metadata records, this information can be used to indirectly rate a current metadata record.	Check the internal metadata store for related entities in the network an calculate a score based on the related data.
19	Usage of digital signatures	Digital signatures can be used to validate a data provider and therefore they can provide a trust measurement.	Check the linked sources for signatures and provide a boolean value for the existence.
Community			
20	Communication	Comprehensible discussions about open data records can raise the awareness of the contents context. Furthermore, online communication tools can actively used to gather additional information about data.	Check, if there are sources of online discussions like message boards or mailing lists given.
21	Trust transfer	A trust value based on user votes can provide insights of others opinions.	Check, if the metadata sources provide a user-based trust ranking and provide the respective information.
22	Correctness	A correctness value based on user votes can provide insights of others opinions.	Check, if the metadata sources provide a user-based correctness ranking and provide

			the respective information.
23	Confirmation	Metadata may be provided by different sources. An overlapping of the same metadata values can be a confirmation, if the sources do not rely on each other.	Check, if several metadata sources provide the same metadata information and calculate a score out of overlappings. If metadata fields do not overlap, the independent fields can be aggregated.
Versatility			
24	Multiple serializations*	Metadata can be provided in different formats, e.g. XML, JSON, or RDF. Especially for mobility data, e.g. Web Map Service (WMS) and Web Feature Service (WFS) could be provided.	Check, if links to different file formats are provided.
25	Multiple languages	Metadata can be provided in different languages, especially unstructured parts like description texts.	Check, if links with different languages in URL or title of the links are provided.
26	Multiple access methods	The same metadata can be provided by file-based repositories, APIs, query languages like SPARQL, or protocols like HTTP or FTP.	Check, if links with different access methods are provided.
Representation			
27	Open format	Metadata should be provided using an open, non-proprietary standard.	Check, if the provided metadata format uses standards from W3C or similar institutions.
28	Data format	The provided format of data should be well-defined, published as an official standard, and be registered at institutions like IANA to support interoperability and access.	Check, if the metadata format is conform with well-known standards.
29	Machine processable	To provide automatic processing, metadata should be structured accordingly.	Check, if the metadata format supports a structure to be automatically processed by machines.
30	Vocabulary	Metadata should be described using a well-structured, well-defined vocabulary/ontology/schema to be comparable, reusable, and automatically processable.	Check, if the used schema for metadata representation is given.
31	Date Format	Dates should be given in a standardized form to avoid parsing errors and issues with time differences.	Check the provided date formats.
32	Unique identifier	To identify a metadata record, an universal unique persistent identifier should be used.	Check, if information about a UID is given.
33	Locality*	Especially for mobility data, information about the geographical context is important. This hierarchical information often can be refined, e.g. if the locality information value is a state which can hierarchically be refined to match also contained regions like cities.	Check, if there is a metadata field for the geographical region or if this information can be extracted out of unstructured texts inside the metadata.
Interlinking			
34	Labeled data	Metadata fields and values inside the fields are labeled to provide at least semi-structured data. Labeled entities should be human and machine readable.	Check, if the data is at least semi-structured (e.g. by using XML) or structured (e.g. by using RDF)

35	Linked data representation	Metadata is provided as linked data and represented as RDF or approaches on higher levels.	Check use of RDF
36	Metadata interlinking	The used metadata representation is extended with links to other metadata standards to support automatic processing and reasoning by machines.	Check, if ontologies/namespaces/schemas of other standards are provided.
Contactability			
37	Contact URL	A contact website of the data provider allows to get in contact and to find related datasets of the same provider.	Check, if a URL of the provider is given.
38	Contact Email	Metadata records can contain a contact email address to reach responsible persons of the data and open up the option for data improvements.	Check, if a contact email is provided and the format is valid (e.g. used characters, use of one at-character, given domain)
39	Classical contact information*	Classical, non-digital contact information like address, telephone number, responsible department or person provide additional options to contact the data provider.	Check, if the metadata contains respective data, ideally structured by fields for adress, telephone number, and responsive actor.
Access			
40	Open metadata	The metadata can be accessed without any restrictions or the need for a registration.	Check, if there is a way to access the metadata in an open way.
41	Retrievability	The metadata should be retrieved by an agent and the response should return a code for success (e.g. HTTP 200 or FTP 2xx)	Check, if the metadata can be accessed and a success code is returned.
Versioning*			
42	Version numbering*	For different states of data, a version number can be given in the metadata to provide an identificator of the current version.	Check, if metadata contains a dedicated field for the version or try to extract it out of description texts inside the metadata.
43	Period of time*	The collection, generation or aggregation of data takes place at a specific period of time. This can be a valuable information for data users.	Check, if a related field and value is part of the metadata record. Additionally, unstructured texts can be checked for this information.
Data			
44	Open data format	Data should be provided using an open, non-proprietary standard. This can be checked by information given in the metadata.	Check, if the provided data format uses standards from W3C or similar institutions.
45	Data format	The provided format of data should be well-defined, published as an official standard, and be registered at institutions like IANA to support interoperability and access. This can be checked by information given in the metadata.	Check, if the data format is conform with well-known standards.
46	Machine processable data	To provide automatic processing, data should be structured accordingly. This can be checked by information given in the metadata and is a specialization of the more general view on data formats.	Check, if the data format supports a structure to be automatically processed by machines.

47	Unique data identifier	To identify a data record, an universal unique persistent identifier should be used. This information can be given in the metadata.	Check, if information about a data UID is given.
48	Multiple data serializations	Data can be provided in different formats, e.g. XML, JSON, or RDF. These can be linked inside the metadata.	Check, if links to different file formats are provided.
*Extensions of the literature review by the data-driven-approach			

Table 0: Aggregated OPAL metadata quality criteria and metrics for mobility data based on the literature review and the data-driven investigation

6 Conclusions

The identified metadata properties for a automatic generation and evaluation of metadata quality comprise 13 dimensions and 48 criteria. The criteria catalogue will be used to implement a quality validation component for the OPAL portal (working package 8).

Appendix

A.A Aggregated quality criteria from literature

No.	Criterion	Description	Metric
Expressiveness			
1	Extend	A metadata entry for a specific dataset should consist of an extensive, non-empty set of metadata fields.	Count the number of given metadata values and divide it by the total number of used metadata fields.
2	Weighted extend	For different fields of interest, specific metadata fields can become a related importance. This can be quantified by applying (user-based) weights for individual metadata fields.	Multiply given metadata fields with a weight, compute the sum of all weighted vales, and divide it of the largest possible value of a sum.
3	Categorization	Especially categories and tags can be used to discover new datasets and vice versa a current dataset can be discovered form other sources.	This can be implemented by checking a true/false value, if a metadata record contains tags/categories or not. Additionally, the number of categories and tags can be used for quantification.
Temporal			
4	Timeliness	Data users often are intersted in up-to-date data.	Calculate the difference of the current time and the time of creation or the last update. Only one timestamp in the metadata is needed.
5	Update rate	Information about the frequency of updates.	Sum up the number of update events.
Understandability			
6	Readability	The readability of text can be estimated by sentence length and parts of speech.	Usage of a readability test, e.g. Flesch-Reading-Ease
7	Language errors	Correct language in metadata fields is can be an indicator for data quality, especially if parts of the metadata have been extracted from the original data.	Count the number of words written incorrectly. Additionally, heuristics for grammar errors can be applied.
8	Example applications	Example applications, which are based on data described by a metadata record can be software projects or articles.	Check, if there are links to examples.
Rights			
9	Machine readable license	A machine readable license can be processed automatically to provide and aggregate information for users.	Check, if there is an ID, URL, or structured information about the license.
10	Human readable license	In user interfaces, listed licenses can provide users with information about rights or at least provide the option to look up the related license permissions and restrictions.	Check, if the metadata contains a title of the license.
11	Known License	Properties of well-known licenses can be represented in different ways to users and are an added value.	Check, if the license is listed in a license database.

12	Open License	Open licenses typically extend the range of possible applications for work with the related data.	Check, if the license is described as open.
13	Permission for commercial use	Information extracted out of data is valuable and can be integrated in commercial databases, applications or just sold to reward the work with data.	Check, if the license is described as free for commercial use.
14	Permissions	Some datasets are restricted to give attribution to the original author. The awareness about these restrictions are important for re-publishing data.	Check, if the license meets properties like attribution, share-alike, or public-domain.
Trust			
15	Provider identity	Data can be provided from instances like government, universities, or arbitrary humans or machines. A known identity can be used for researching additional information	Check, if the official name, title, or website of the data provider is provided.
16	Trusted provider	The credibility and reliability of the metadata provider can be stored in a database and metadata records from the same provider can be rated.	Check, if the metadata provider is listed in a provider database with rating information.
17	Metadata authenticity	If several publishers or sources for one metadata record are found or if there are overlappings among data contributors for several metadata records, this information can be used to indirectly rate a current metadata record.	Check the internal metadata store for related entities in the network an calculate a score based on the related data.
18	Usage of digital signatures	Digital signatures can be used to validate a data provider and therefore they can provide a trust measurement.	Check the linked sources for signatures and provide a boolean value for the existance.
Community			
19	Communication	Comprehensible discussions about open data records can raise the awareness of the contents context. Furthermore, online communication tools can actively used to gather additional information about data.	Check, if there are sources of online discussions like message boards or mailing lists given.
20	Trust transfer	A trust value based on user votes can provide insights of others opinions.	Check, if the metadata sources provide a user-based trust ranking and provide the respective information.
21	Correctness	A correctness value based on user votes can provide insights of others opinions.	Check, if the metadata sources provide a user-based correctness ranking and provide the respective information.
22	Confirmation	Metadata may be provided by different sources. An overlapping of the same metadata values can be a confirmation, if the sources do not rely on each other.	Check, if several metadata sources provide the same metadata information and calculate a score out of overlappings. If metadata fields do not overlap, the independent fields can be aggregated.
Versatility			
23	Multiple serializations	Metadata can be provided in different formats, e.g. XML, JSON, or RDF	Check, if links to different file formats are provided.

24	Multiple languages	Metadata can be provided in different languages, especially unstructured parts like description texts.	Check, if links with different languages in URL or title of the links are provided.
25	Multiple access methods	The same metadata can be provided by file-based repositories, APIs, query languages like SPARQL, or protocols like HTTP or FTP.	Check, if links with different access methods are provided.
Representation			
26	Open format	Metadata should be provided using an open, non-proprietary standard.	Check, if the provided metadata format uses standards from W3C or similar institutions.
27	Data format	The provided format of data should be well-defined, published as an official standard, and be registered at institutions like IANA to support interoperability and access.	Check, if the metadata format is conform with well-known standards.
28	Machine processable	To provide automatic processing, metadata should be structured accordingly.	Check, if the metadata format supports a structure to be automatically processed by machines.
29	Vocabulary	Metadata should be described using a well-structured, well-defined vocabulary/ontology/schema to be comparable, reusable, and automatically processable.	Check, if the used schema for metadata representation is given.
30	Date Format	Dates should be given in a standardized form to avoid parsing errors an issues with time differences.	Check the provided date formats.
31	Unique identifier	To identify a metadata record, an universal unique persistent identifier should be used.	Check, if information about a UID is given.
Interlinking			
32	Labeled data	Metadata fields an values inside the fields are labeled to provide at least semi-structured data. Labeled entities should be human and machine readable.	Check, if the data is at least semi-structured (e.g. by using XML) or structured (e.g. by using RDF)
33	Linked data representation	Metada is provided as linked data and represented as RDF or approaches on higher levels.	Check use of RDF
34	Metadata interlinking	The used metadata representation is extended with links to other metadata standards to support automatic processing and reasoning by machines.	Check, if ontologies/namespaces/schemas of other standards are provided.
Contactability			
35	Contact URL	A contact website of the data provider allows to get in contact and to find related datasets of the same provider.	Check, if a URL of the provider is given.
36	Contact Email	Metadata records can contain a contact email adress to reach responisble persons of the data and open up the option for data improvements.	Check, if a contact email is provided and the format is valid (e.g. used characters, use of one at-character, given domain)

Access			
37	Open metadata	The metadata can be accessed without any restrictions or the need for a registration.	Check, if there is a way to access the metadata in an open way.
38	Retrievability	The metadata should be retrieved by an agent and the response should return a code for success (e.g. HTTP 200 or FTP 2xx)	Check, if the metadata can be accessed and a success code is returned.
Data			
39	Open data format	Data should be provided using an open, non-proprietary standard. This can be checked by information given in the metadata.	Check, if the provided data format uses standards from W3C or similar institutions.
40	Data format	The provided format of data should be well-defined, published as an official standard, and be registered at institutions like IANA to support interoperability and access. This can be checked by information given in the metadata.	Check, if the data format is conform with well-known standards.
41	Machine processable data	To provide automatic processing, data should be structured accordingly. This can be checked by information given in the metadata and is a specialization of the more general view on data formats.	Check, if the data format supports a structure to be automatically processed by machines.
42	Unique data identifier	To identify a data record, an universal unique persistent identifier should be used. This information can be given in the metadata.	Check, if information about a data UID is given.
43	Multiple data serializations	Data can be provided in different formats, e.g. XML, JSON, or RDF. These can be linked inside the metadata.	Check, if links to different file formats are provided.

Table A: Aggregated quality criteria from literature

A.G 8 Principles of Open Government Data

No.	Criterion	Description
1	Complete	All public data is made available. Public data is data that is not subject to valid privacy, security or privilege limitations.
2	Primary	Data is as collected at the source, with the highest possible level of granularity, not in aggregate or modified forms.
3	Timely	Data is made available as quickly as necessary to preserve the value of the data.
4	Accessible	Data is available to the widest range of users for the widest range of purposes.
5	Machine processable	Data is reasonably structured to allow automated processing.
6	Non-discriminatory	Data is available to anyone, with no requirement of registration.
7	Non-proprietary	Data is available in a format over which no entity has exclusive control.
8	License-free	Data is not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed.
Aaron Swartz et al. (2007). 8 Principles of Open Government Data. http://web.archive.org/web/20071214024243/https://public.resource.org/8_principles.html		

Table G: Results from Swartz et al. (2007)

A.L Linked Data - Design Issues

No.	Criterion	Description
1	OpenLicence	Available on the web (whatever format) but with an open licence, to be Open Data
2	MachineRead	Available as machine-readable structured data (e.g. excel instead of image scan of a table)
3	OpenFormat	Available in non-proprietary format (e.g. CSV instead of excel)
4	OpenStandard	Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
5	LinkedData	Link your data to other people's data to provide context
Berners-Lee, T. (2010). Linked Data - Design Issues. http://web.archive.org/web/20101202183255/https://www.w3.org/DesignIssues/LinkedData.html		

Table L: Results from Berners-Lee (2010)

A.Z Quality assessment for Linked Data

A.Z.A Quality assessment for Linked Data: Accessibility dimensions

No.	Criterion	Description
Availability		
1	accessibility of the SPARQL endpoint and the server	checking whether the server responds to a SPARQL query
2	accessibility of the RDF dumps	checking whether a RDF dump is provided and can be downloaded
3	dereferencability issues	when a URI returns an error (4xx client error/ 5xx server error) response code or detection of broken links
4	no structured data available	detection of dead links or detection of a URI without any supporting RDF metadata or no redirection using the status code 303 See Other or no code 200 OK
5	no dereferenced back-links	detection of all local in-links or back-links: locally available triples in which the resource URI appears as an object, in the dereferenced document returned for the given resource
6	no dereferenced forward-links	detection of all forward links: locally known triples where the local URI is mentioned in the subject
7	misreported content types	detection of whether the content is suitable for consumption, and whether the content should be accessed
Licensing		
8	machine-readable indication of a license	detection of the indication of a license in the VoID description or in the dataset itself
9	human-readable indication of a license	detection of a license in the documentation of the dataset or its source
10	permissions to use the dataset	detection of license indicating whether reproduction, distribution, modification or redistribution is permitted
11	indication of attribution, Copyleft or ShareAlike	detection of whether the work is attributed in the same way as specified by the author or licensor
Interlinking		

12	interlinking degree, clustering coefficient, centrality and sameAs chains, description richness through sameAs	by using network measures
13	existence of links to external data providers	detection of the existence and usage of external URIs and owl:sameAs links
Security		
14	access to data is secure	use of login credentials or use of SSL or SSH
15	data is of proprietary nature	data owner allows access only to certain users
Performance		
16	no usage of slash-URIs	checking for usage of slash-URIs where large amounts of data is provided
17	low latency	delay between submission of a request by the user and reception of the response from the system
18	high throughput	no. of answered HTTP-requests per second
19	scalability of a data source	detection of whether the time to answer an amount of ten requests divided by ten is not longer than the time it takes to answer one request
Accessibility dimensions Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for Linked Data: A Survey: A systematic literature review and conceptual framework. Semantic Web, 7(1)		

Table Z.A: Results from Zaveri et al. (2016)

A.Z.I Quality assessment for Linked Data: Intrinsic dimensions

No.	Criterion	Description
Accuracy		
1	detection of outliers	by using distance-based, deviations-based and distribution-based method
2	inaccurate values	by using functional dependencies rules between the values of two or more different predicates
3	inaccurate facts	a single fact is checked individually in different datasets
4	malformed datatype literals	detection of ill-typed literals which do not abide by the lexical syntax for their respective datatype
5	literals incompatible with datatype range	detection of a datatype clash that can then occur if the property is given a value (i) that is malformed, or (ii) that is a member of an incompatible datatype
6	erroneous annotation/ representation erroneous	1 – (erroneous instances total / no. of instances inaccurate instances)
7	inaccurate annotation, labelling, classification	1 – (inaccurate instances / total no. of instances) * (balanced distance metric / total no. of instances)
Consistency		
8	entities as members of disjoint classes	(no. of entities described as members of disjoint classes / total no. of entities described in the dataset)
9	usage of homogeneous datatypes	no. of properties used with homogeneous units in the dataset / total no. of properties used in the dataset

10	invalid usage of undefined classes and properties	detection of classes and properties used without any formal definition
11	misplaced classes or properties	using entailment rules that indicate the position of a term in a triple
12	misuse of owl:datatypeProperty or owl:objectProperty	by using weighting scheme that identifies that most usage is contrary to the vocabulary constraint
13	use of members of owl:DeprecatedClass or owl:-DeprecatedProperty	based on a manual mapping between deprecated terms and compatible term
14	provide a blacklist for void values	list all bogus owl:Inverse-FunctionalProperty values
15	ontology hijacking	detection of the redefinition by third parties of external classes/ properties such that reasoning over data using those external terms is affected
16	misuse of predicates	profiling statistics support the detection of such discordant values or misused predicates and facilitate to find valid formats for specific predicates
17	ambiguous annotation	1 - (no. of ambiguous instances / no. of the instances contained in the semantic metadata set)
Interlinking		
18	intensional conciseness	no. of unique attributes of a dataset / total no. of attributes in a target schema
19	extensional conciseness	no. of unique objects of a dataset / total number of objects representations in the dataset
20	duplicate instance	1 – (total no. of instances that violate the uniqueness rule / total no. of relevant instances)
Intrinsic dimensions Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for Linked Data: A Survey: A systematic literature review and conceptual framework. Semantic Web, 7(1)		

Table Z.I: Results from Zaveri et al. (2016)

A.Z.T Quality assessment for Linked Data: Trust dimensions

No.	Criterion	Description
1	detection of outliers	by assigning explicit ratings to the dataset (manual) and analyzing external links or page rank (semi-automated)
2	meta-information about the identity of information provider	checking whether the provider/contributor is contained in a list of trusted providers
3	indication of metadata about a dataset (provenance information)	presence of the title, content and URI of the dataset
4	computing the trustworthiness of RDF statements	computing a trust value based on the provenance information which can be either unknown or a value in the interval [-1,1] where 1: absolute belief, -1: absolute disbelief and 0:lack of belief/disbelief
5	computing the trust of an entity	construction of decision networks informed by provenance graphs

6	accuracy of computing the trust between two entities	by using a combination of (1) a propagation algorithm which utilizes statistical techniques for computing trust values between 2 entities through a path and (2) an aggregation algorithm based on a weighting mechanism for calculating the aggregate value of trust over all paths
7	acquiring content trust from users	based on associations that transfer trust from entities to resources
8	assigning trust values to data/sources/rules	use of trust ontologies that assign content-based or metadata-based trust values that can be transferred from known to unknown data
9	determining trust value for data	using annotations for data such as (i) blacklisting, (ii) authoritativeness and (iii) ranking and using reasoning to incorporate trust values to the data
10	computing personalized trust recommendations	using provenance of existing trust annotations in social networks
11	detection of reliability and credibility of a data source	use of trust annotations made by several individuals to derive an assessment of the sources' reliability and credibility
12	computing the trustworthiness of RDF statements	computing a trust value based on user-based ratings or opinion-based method
13	detect the reliability and credibility of the dataset publisher	indication of the level of trust for the publisher on a scale of 1 – 9
14	authenticity of the dataset	verifying authenticity of the dataset based on a provenance vocabulary such as the author and his contributors, the publisher of the data and its sources if any
15	usage of digital signatures	by signing a document containing an RDF serialization or signing an RDF graph
16	correctness of the dataset	verifying correctness of the dataset with the help of unbiased trusted third party
17	objectivity of the information	checking for bias or opinion expressed when a dataprovider interprets or analyzes facts
18	objectivity of the source	checking whether independent sources confirm a fact
19	no biased data provided by the publisher	checking whether the dataset is neutral or the publisher has a personal influence on the data provided
Intrinsic dimensions Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for Linked Data: A Survey: A systematic literature review and conceptual framework. Semantic Web, 7(1)		

Table Z.T: Results from Zaveri et al. (2016)

A.Z.D Quality assessment for Linked Data: Dynamicity dimensions

No.	Criterion	Description
Currency		
1	currency of documents/statements	1 – (observation time – last modified time) / (observation time – publishing time)
2	time since modification	observation time - last modified time
3	exclusion of outdated data	1 – (outdated data / total amount of data)
Volatility		
4	frequency of change	refer to the changefrequency attribute in a Semantic Sitemap for value of the frequency or updates of a data source
5	time validity interval	expiry time – input time of the semantic web source
Timeliness		
6	timeliness between the semantic source web and original source	a positive difference between last modified time of the original source and last modified time of the semantic web source implies data source to be outdated
7	timeliness of the resource	a positive difference between current and expiry time of the resource implies data source to be outdated
8	timeliness between the ideal freshness and the data source freshness	1 – (observation time – last modified time) / ideal freshness

Dynamicity dimensions
Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for Linked Data: A Survey: A systematic literature review and conceptual framework. Semantic Web, 7(1)

Table Z.D: Results from Zaveri et al. (2016)

A.Z.C Quality assessment for Linked Data: Contextual dimensions

No.	Criterion	Description
Completeness		
1	schema completeness	no. of classes and properties represented / total no. of classes and properties
2	property completeness	no. of values represented for a specific property / total no. of values for a specific property
3	population completeness	no. of real-world objects are represented / total no. of real-world objects
4	interlinking completeness	no. of instances in the dataset that are interlinked / total no. of instances in a dataset
Amount-of-data		
5	appropriate volume of data for a particular task	ratio of no. of semantically valid association rules to the no. of non-trivial rules
6	appropriate amount of data	use of the apriori algorithm to detect poor predicates based on the occurrence dependencies among predicates
7	amount of triples	no. of triples present in a dataset
8	coverage	scope (no. of entities) and level of detail (no. of properties)
Relevancy		
9	usage of meta-information attributes	counting the occurrence of relevant terms within these attributes or using vector space model and assigning higher weight to terms that appear within the meta-information attributes
10	retrieval of relevant resources	sorting documents according to their relevancy for a given query
Contextual dimensions Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for Linked Data: A Survey: A systematic literature review and conceptual framework. Semantic Web, 7(1)		

Table Z.C: Results from Zaveri et al. (2016)

A.Z.R Quality assessment for Linked Data: Representational dimensions

No.	Criterion	Description
Representational-conciseness		
1	keeping URIs short	detection of long URIs or those that contain query parameters
2	no use of prolix RDF features	detect use of RDF primitives i.e. RDF reification, RDF containers and RDF collections
Representational-consistency		
3	re-use existing terms	detect whether existing terms from other vocabularies have been reused
4	re-use existing vocabularies	usage of established vocabularies
Understandability		
5	human-readable labelling of classes, properties and entities	percentage of entities having an rdfs:label or rdfs:comment
6	dereferenced representations: providing human-readable metadata	detecting the use of rdfs:label to attach labels or names to resources

7	indication of one or more exemplary URIs	detecting whether the pattern of the URIs is provided
8	indication of a regular expression that matches the URIs of a dataset	detecting whether a regular expression that matches the URIs is present
9	indication of an exemplary SPARQL query	detecting whether examples of SPARQL queries are provided
10	indication of the vocabularies used in the dataset	checking whether a list of vocabularies used in the dataset is provided
11	provision of message boards and mailing lists	checking the effectiveness and the efficiency of the usage of the mailing list and/or the message boards
Interpretability		
12	use of self-descriptive formats	identifying objects and terms used to define these objects with globally unique identifiers
13	interpretability of terms	use of various schema languages to provide definitions for terms
14	interpretability of data	detect the use of appropriate language, symbols, units and clear definitions
15	misinterpretation of missing values	detecting use of blank nodes
16	atypical use of collections, containers and reification	detect non-standard usage of collections, containers and reification (since these features are discouraged from use by Linked Data guidelines)
Versatility		
17	provision of the data in different serialization formats	checking whether data is available in different serialization formats
18	provision of the data in various languages	checking whether data is available in different languages
19	accessing of data in different ways	checking whether the data is available as SPARQL endpoint and for download as RDF dump
20	application of content negotiation	checking whether data can be retrieved in accepted formats and languages by adding a corresponding accept-header to an HTTP request
Representational dimensions Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for Linked Data: A Survey: A systematic literature review and conceptual framework. Semantic Web, 7(1)		

Table Z.R: Results from Zaveri et al. (2016)

A.W The FAIR Guiding Principles

No.	Criterion	Description
Findable		
1	F1	(meta)data are assigned a globally unique and persistent identifier
2	F2	data are described with rich metadata (defined by R1 below)
3	F3	metadata clearly and explicitly include the identifier of the data it describes
4	F4	(meta)data are registered or indexed in a searchable resource
Accessible		
5	A1	(meta)data are retrievable by their identifier using a standardized communications protocol
6	A1.1	the protocol is open, free, and universally implementable
7	A1.2	the protocol allows for an authentication and authorization procedure, where necessary
8	A2	metadata are accessible, even when the data are no longer available
Interoperable		
9	I1	(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
10	I2	(meta)data use vocabularies that follow FAIR principles
11	I3	(meta)data include qualified references to other (meta)data
Reusable		
12	R1	meta(data) are richly described with a plurality of accurate and relevant attributes
13	R1.1	(meta)data are released with a clear and accessible data usage license
14	R1.2	(meta)data are associated with detailed provenance
15	R1.3	(meta)data meet domain-relevant community standards

Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3. https://doi.org/10.1038/sdata.2016.18

Table W: Results from Wilkinson et al. (2016)

A.R Assessment and Visualization of Metadata Quality for Open Government Data

No.	Criterion	Description
Existence		
1	Completeness	A metadata record is considered complete, if the record contains all the information requiredhave an ideal representation of the described resource. [...]
2	Weighted Completeness	While the completeness metric is straightforward it comes with the drawback of treating every field with the same importance. The relevance of a certain metadata field depends strongly on the context. [...]
3	Accuracy	The accuracy metric measures how accurate the metadata record represents the associated resources. [...]
4	Richness of Information	Richness of Information The vocabulary terms and the description used in a metadata record should be meaningful to the user. For that the metadata need to contain enough information for describing uniquely the referred resource. [...]
5	Readability	The readability metric measures the degree to which a metadata record is cognitive accessible. The readability describes how easy a user can comprehend what the resource is about after reading the metadata record. [...]
6	Availability	Metadata records contain URLs which point to the actual resources. assesses the number of reachable resources. [...]
7	Misspelling	Readers which are proficient in a language might halt for a moment on words written incorrectly. [...]
Reiche, K. J., Höfig, E., & Schieferdecker, I. (2014). Assessment and Visualization of Metadata Quality for Open Government Data. In Conference for E-Democracy and Open Governement (pp. 335--346). Krems.		

Table R: Results from Reiche et al. (2014)

A.U Quality Assessment and Evolution of Open Data Portals

No.	Criterion	Description
Existence		
1	Retrievability	The extent to which meta data and resources can be retrieved.
2	Usage	The extent to which available meta data keys are used to describe a dataset.
3	Completeness	The extent to which the used meta data keys are non empty.
4	Accuracy	The extent to which certain meta data values accurately describe the resources.
5	Openness	The extent to which licenses and file formats conform to the open definition.
6	Contactability	The extent to which the data publisher provide contact information.
Umbrich, J., Neumaier, S., & Polleres, A. (2015). Quality Assessment and Evolution of Open Data Portals (pp. 404–411). IEEE. https://doi.org/10.1109/FiCloud.2015.82		

Table U: Results from Umbrich et al. (2015)

A.N Automated Quality Assessment of Metadata across Open Data Portals

No.	Criterion	Description
Existence		
1	Access	Is there access information for resources provided?
2	Discovery	Is information available that can help to discover/search datasets?
3	Contact	Existence of information that would allow to contact the dataset provider.
4	Rights	Existence of information about the license of the dataset or resource.
5	Preservation	Existence of information about format, size or update frequency of the resources
6	Date	Existence of information about creation and modification date of metadata and resources respectively.
Conformance		
7	AccessURL	Are the values of access properties valid HTTP URLs?
8	ContactEmail	Are the values of contact properties valid emails?
9	ContactURL	Are the values of contact properties valid HTTP URLs?
10	DateFormat	Is date information specified in a valid date format?
11	Licence	Can the license be mapped to the list of licenses reviewed by opendefinition.org?
12	FileFormat	Is the specified file format or media type registered by IANA?
Retrievability		
13	Retrievable	Can the described resources be retrieved by an agent?
Accuracy		
14	FormatAccuracy	Is the specified file format accurate?
15	SizeAccuracy	Is the specified file size accurate?
Open Data		
16	OpenFormat	Is the file format based on an open standard?
17	MachineReadable	Can the file format be considered as machine readable?
18	OpenLicense	Is the used license conform to the open definition?
Neumaier, S., Umbrich, J., & Polleres, A. (2016). Automated Quality Assessment of Metadata across Open Data Portals. Journal of Data and Information Quality, 8(1), 1–29. https://doi.org/10.1145/2964909		

Table N: Results from Neumaier et al. (2016)