A Schema Readability Metric for Automated Data Quality Measurement

Lisa Ehrlinger*†, Gudrun Huszar*, Wolfram Wöß*
*Johannes Kepler University Linz, Altenberger Straße 69, 4040 Linz, Austria
†Software Competence Center Hagenberg, Softwarepark 21, 4232 Hagenberg, Austria
email: lisa.ehrlinger@jku.at, wolfram.woess@jku.at

***Abstract*—Data quality measurement is a critical success factor to estimate the explanatory power of data-driven decisions. Several data quality dimensions, such as completeness, accuracy, and timeliness, have been investigated so far and metrics for their measurement have been proposed. While most research into those dimensions refers to the data values, schema quality dimensions in general, and readability in particular, have not gained sufficient attention so far. A poorly readable schema has a negative impact on the data quality, e.g., two attributes with different purpose, but synonymous labels may cause incorrectly inserted attribute values. Thus, we specifically observe the data quality dimension *readability* on schema-level and introduce a metric for its measurement. The measurement is based on a dictionary-approach using a wordnet, which takes into account the semantics of the words used in the schema (e.g., attribute labels). We implemented and evaluated the schema readability metric within the data quality tool QuaIIe.**

***Index Terms*—Data Quality; Metrics; Readability; Semantics.**

I. Introduction

Data Quality (DQ) is a prerequisite to trust data-driven decisions, which can, for example, be strategic decisions in companies, or artificial intelligence algorithms for self-driving cars. Eckerson [1] estimated the costs arising from poor customer data for companies to be more than 600 billion US dollars a year. These costs include failed prints and loss of customers due to incorrect addressing, as well as staff overhead. According to Loshin [2], the primary categories of negative impacts related to DQ are financial (e.g., decreased revenues and increased penalties), confidence and satisfaction-based impacts, productivity impacts (e.g., decreased throughput), and risk and compliance impacts (e.g., investment risks).

Data quality is usually measured in different dimensions, such as, completeness, accuracy, consistency, and minimality [3][4]. Those dimensions can either refer to the extension of the data (i.e., data values), or to their intension (i.e., the schema) [4]. While a lot of research has been conducted for DQ dimensions on the data-level (cf. [5]–[8]), schema quality dimensions in general, and readability in particular, have not gained sufficient attention so far. In existing research, the measurement of readability is usually associated with textual documents and not primarily to Information Systems (ISs). To the best of our knowledge, there exists no metric to measure the readability of IS schemas. Thus, the major contribution of this paper is a discussion of the schema quality dimension readability along with a newly developed metric for its measurement. An essential feature of the metric is the incorporation of semantics of attribute labels using a wordnet.

According to Vossen [9], the quality dimension *readability* describes the condition, in which a schema represents the modeled domain in a natural and clear way, which means, it is self-explanatory to the user. From a more general perspective, the readability of IS schemas is important for two aspects: (1) the understandability of a schema for humans, as described by [9], and (2) the degree to which a schema can be used for automated schema fusion, integration, or matching approaches. An example are two IS schemas within a company, where one schema has a table `product` for storing product types, and the second schema has a corresponding table `prod.Type`, which stores the same entity type. An automated schema integration algorithm requires a sufficient level of readability and standardization in order to merge both tables. Also, an employee, who is not familiar with the schemas, might consider the tables as not equivalent. This scenario could lead on the one hand to duplicate entity types (because both tables are populated separately), and on the other hand to incomplete inventory counts (because only one table is queried for sales statistics). To show the applicability of our readability metric, we implemented it in the DQ tool QuaIIe [10] and evaluated the ratings of several databases (DBs).

This paper is structured as follows: Section II summarizes related work concerning the measurement of *readability*. In Section III, we present our approach how to measure the readability of IS schemas, with our newly developed metric. The metric is demonstrated and discussed in Section IV. We conclude in Section V with an outlook on future work.

II. State of the Art and Related Work

In this section, we provide an overview of related work about readability and explain why existing readability metrics are not sufficiently developed. The DQ dimension readability is most commonly described as the degree to which a schema represents the modeled domain in a natural and clear way, with the aim to be self-explanatory to the user [9]. Since clarity is subjective, no generally valid formal definition for this DQ dimension exists [4]. In alignment with the “fitness for use” principle of DQ [3][11], the readability dimension depends on the intended use and user group. For this definition, the knowledge of the user, the vocabulary and the format of the data is important. In addition to the user perspective, readability is an important aspect for automated schema matching approaches, e.g., in the area of information fusion or information fusion.

When considering the topic from a more general viewpoint, research about the readability of texts in documents has already been published since 1900 [12][13]. In those philology-based approaches, sentence features (e.g., sentence length, syllables in words, word length and popularity) are used to measure the readability of texts. Renzis et al. [14] define readability in this context as the difficulty or simplicity of text comprehension for the intended user. One frequently used index is the Automated Readability Index (ARI) [13], which computes the readability based on syllables per word:

$$ARI = \frac{w}{s} + 9 * \frac{z}{w}, \qquad (1)$$

where $\frac{w}{s}$ is the number of words w per sentence s, and $\frac{z}{w}$ is the number of characters z per word. However, Zhao and Khan [12] observed that such philology-based approaches do not consider domain-specific terms sufficiently. For example, *myocardium* is shorter than *myocardal muscle* and thus easier to read with respect to the philology-based approach. The words are synonyms, but a non-expert will rate texts with *myocardium* less readable than texts with *myocardal muscle*. Consequently, readability measures might not represent the "real" readability for non-experts adequately, if domain-specific terms are not considered [12].

However, those philology-based approaches are not useful for measuring schema readability, because instead of sentences, only single words (e.g., attribute labels) are available. While the readability of a conceptual schema in its graphical representation also includes aesthetic criteria, such as the arrangement of entities or crossing lines [15], the readability of a logical schema is limited to the actual naming of entities and relationships.

In the frame of DQ research, Cai et al. [5] observed DQ standards for big data in five dimensions including *presentation quality*, which covers the readability and structure of data representation. Data with high presentation quality allows the user to understand and interpret the data. However, this understanding requires knowledge about commonly used terms, for example, units, codes, and abbreviations. Cai et al. [5] suggested the following indicators to assess the degree of readability in data: (a) data (content, format, semantics, etc.) are clear and understandable, (b) it is easy to judge that the data provided meet requirements, and (c) data description, classification, and coding content satisfy specification and are easy to understand. There is no formal definition of those three indicators, which would allow a direct application to measure the readability in a company IS. The first indicator *clear and understandable* [5] is closely connected with the term *comprehension* from the philological readability definition by Renzis et al. [14]. If a text is clear and understandable, a reader can simply comprehend it. In the context of data, a human can interpret the data and eventually derive information and knowledge.

Yan et al. [16] presented a domain-specific and ontology-based readability measure, which is based on two document properties: cohesion and scope. Cohesion refers to the relatedness of words and is influenced by the association of terms in an ontology. The closer the words in an ontology, the higher is the cohesion. The scope refers to experts knowledge. Assuming $n > 1$, and $i < j$, cohesion is calculated according to [16]:

$$Cohesion(d_i) = \frac{\sum_{i,j=1}^{n} Sim(c_i, c_j)}{NumberOfAssociations}, \qquad (2)$$

$$Sim(c_i, c_j) = -log\frac{len(c_i, c_j)}{2D}, \qquad (3)$$

$$NumberOfAssociations = \frac{n(n-1)}{2}, \qquad (4)$$

where d_i is a document, *n* is the total number of domain concepts, and *c* is a concept. *Sim*(c_i, c_j) computes semantic similarity of concepts. The function *len*(c_i, c_j) calculates the shortest path between two concepts. *NumberOfAssociations* is the total number of associations among domain concepts.

All mentioned approaches do not provide a definition nor a metric for readability of IS schemas. Thus, we tackle this research issue with a specification of the DQ dimension readability on IS schema-level and a metric to measure it, which is presented in the following section. The approach aims at automated readability measurement, which can be employed for continuous DQ monitoring.

III. An Approach to Measure Schema Readability

In this section, we present our approach to achieve a sufficient level of readability in IS schemas. The approach can be divided into three steps, which are explained in the following subsections: (1) schema preprocessing in order to achieve comparability of different schemas and extract words for the readability calculation, (2) the development of a set of readability criteria, and based on these criteria, (3) the calculation of our readability metric.

A. Schema Preprocessing

For each evaluated schema, a machine-readable description using the Data Source Description (DSD) vocabulary [17] is generated. The DSD vocabulary is an abstraction layer for different schemas. An excerpt of such a DSD file is shown in Figure 1, which contains the description of the `employees` table and the attribute "first_name" from the employees DB [18].

The labels (`rdfs:label`) contained in the DSD files are the basis to extract "words" for further processing. A word can be either the complete label, or part of it. If a schema uses delimiters like underscores (_), hyphens (-), or camel case, one label is split into several words. For example, "first_name" (or alternatively "firstName") is split into "first" and "name". This string splitting enables the usage of each substring of a concatenated label for the readability calculation. One IS schema may consist of several concepts, which are, e.g., tables in relational DBs. In that case, the readability is calculated for

```
ex:employees a dsd:Concept;
  rdfs:label          "employees";
  dsd:hasPrimaryKey   ex:employees.pk;
  dsd:hasAttribute    ex:employees.emp_no,
                      ex:employees.first_name,
                      ex:employees.last_name,
                      ex:employees.birth_date,
                      ex:employees.hire_date,
                      ex:employees.gender.

ex:employees.first_name a dsd:Attribute;
  rdfs:label              "first_name";
  dsd:isOfDataType        xsd:string;
  dsd:maxCharacterLength  "14"^^xsd:long ;
  dcterms:title           "first_name" .
```

Fig. 1. Data Source Description of Employees

each concept and the mean of all concept-level readability ratings is used as overall rating for the entire schema.

One challenge faced during this work was the accumulation of duplicate words due to the splitting of concatenated strings. Prefixes and suffixes are a common tool to associate attributes to the respective concepts, e.g., “employeeName” and “employeeNumber”. After the splitting, a large number of duplicates (e.g., in this case “employee”) is generated and needs to be further processed. We resolved this issue in the implementation by storing all words in a hashmap and, thus, those duplicate words are only considered once.

B. *Readability Criteria*

For our approach, we developed a set of readability criteria, which are applied to “words”. In the following paragraphs, each of these criteria is discussed in more detail and exemplified with the help of a DB for storing employees (cf. Table I). As a result of the readability calculation proposed in this paper, a quality report is produced, which in addition to the readability rating (from the metric) contains a set of annotations that provide further information about the quality of a schema. Figure 2 shows a flowchart diagram of our approach, including the extraction of words from a DSD file, the evaluation of the criteria, and the annotations that are set for each criterion.

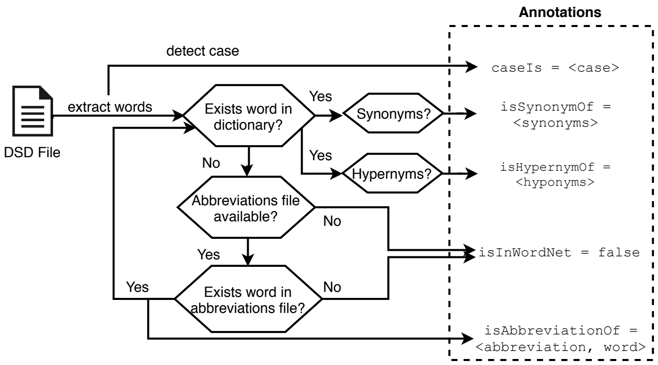
Annotations
detect case
caseIs = <case>
extract words
DSD File
Exists word in dictionary?
Yes
Synonyms?
isSynonymOf = <synonyms>
Yes
Hypernyms?
isHypernymOf = <hyponyms>
No
Abbreviations file available?
No
isInWordNet = false
Yes
Yes
Exists word in abbreviations file?
No
isAbbreviationOf = <abbreviation, word>

Fig. 2. Readability Measurement Approach

1) Wordnet existence: To detect and process cognates (e.g., synonyms and hypernyms), it must be initially checked whether a word exists in a publicly available online dictionary, and therefore can be considered as generally known. For an automated approach, the usage of a wordnet, which is a combination of a dictionary and a thesaurus, is reasonable. A comprehensive list of wordnets is provided in [19] with prominent examples like DBPedia [20], WoNeF [21], or WordNet [22]. For our approach, we selected the widely used WordNet [22][23], which is developed at Princeton University since 1985. In contrast to a dictionary, the terms in WordNet are categorized into nouns, verbs, adjectives, adverbs, and functors, and are sorted according to their semantics [24]. Terms are grouped to sets of synonyms, so called synsets. The structure of WordNet is based on *psycho-linguistics*, which is the science of the human psyche and explores the task of learning and using a language [25]. A word is annotated with *isInWordNet* (set to true or false) to indicate if it is in WordNet.

If a word was not found in the wordnet, it still may be an abbreviation. Abbreviations can impede readability, because they might lead to ambiguities. An example is the abbreviation *MI*, which refers to *Myocardial Infarction* (heart attack) in the medical context, but could also stand for the state Michigan. Furthermore, there exist ambiguities within a single domain. *MI* can, for example, also refer to *Mental Illness*, a mental disorder of a person. Depending on a persons field of expertise (cardiology or neurology), the same abbreviation would be interpreted differently. In our approach, it is possible to add domain-specific abbreviations, which are frequently used in a specific context, but are not contained in WordNet. This measure is also recommended by Hoberman [15] to increase the readability of conceptual IS schemas.

In QualIe, it is possible to add abbreviations in form of a Comma-Separated Values (CSV) file [26]. If such a file is provided and contains a word, which was not found in a wordnet, the annotation *isAbbreviationOf* is set to link the abbreviation to its corresponding full word. An example is provided in Table I, where a relation for storing employees includes an attribute with the label “emp”, which is an abbreviation for “employee”. Without additional information, this label would not be found in a wordnet and no further processing (e.g., checking for synonyms) would be possible.

TABLE I. EMPLOYEES TABLE

emp	worker	SALARY	date	product	ware
Doe	Jones	1400	01012010	car	wheel
Smith	Green	1600	01042018	bike	settle

2) Consistent cases: The consistent use of cases is important for a readable schema [15]. Possible variants are uppercase only, initial uppercase, lowercase, camel case, with or without blanks and/or hyphens. If one attribute is written in lowercase and another attribute in uppercase, this might lead to ambiguities. Thus, the inconsistent usage of cases decreases the readability rating. In addition, the annotation *caseIs* gives evidence about the case detected per word. The attribute "SALARY" in Table I is an example for inconsistently used cases compared to the other attributes.

3) Cognates: The semantics, that is, the meaning of the words, is the most important aspect for humans to interpret, understand, and efficiently work with a IS schema. The term *cognates* has its origin in linguistics and describes related words, which have the same origin or share the same meaning, for example, synonyms, hypernyms, or homonyms. Cognates can lead to ambiguities and therefore to a less readable IS schema. Those relations are considered in our readability metric and discussed in the following paragraphs. Josko et al. [27] defined "synonymous values" and "homonymous values" in their formal taxonomy on data defects on IS content-level. We refined the original definitions from [27] to Definitions 1 and 2, to adopt them to synonyms and homonyms within IS schemas.

a) Synonyms: The term *synonym* is derived from the Greek word "syn", which means "together", and describes words, which share the same meaning.

Definition 1 (Synonyms [27]): Let $sp : w(S) \times w(S) \rightarrow \{true, false\}$ be a function that returns if the graphy and pronunciation of two words within S are equal, according to LEX. Let $me : w(S) \times w(S) \rightarrow \{true, false\}$ be a function that returns if the meaning of two words within S are equal or nearly the same, according to LEX. A schema has synonyms *iff* $\exists w_i, w_j \in S$, where $i \neq j$, such that $sp(w_i, w_j) = false$ and $me(w_i, w_j) = true$. Synonyms denote distinct terms in writing that share the same or similar meanings. Such terms can be expressed as vernacular words, acronyms, abbreviations, or symbols. This defect arises when synonymous terms are used interchangeably to indicate the same fact about objects within a schema.

Here, LEX is a universal thesaurus (i.e., a set of lexical definitions, relationships and similarity degrees [27]) and $w(S)$ the set of n words $\{w_1, w_2.., w_n\}$ within an IS schema S. The attributes "product" and "ware" in Table I are synonyms, because the distinction between the two words is not clear. Thus, the existence of synonyms in an IS schema decreases the readability rating. Additionally, the affected attributes are annotated with *isSynonymOf* to link them to their corresponding synonyms.

b) Hypernyms: The term *hypernym* is derived from the Greek word "hyper", which means "above", and denotes a superordinate concept [25]. An IS schema, which includes a specific word (i.e., a *hyponym*), as well as its superordinate concept (hypernym), leads to ambiguities in the interpretation of a schema. Thus, we decrease the readability, if hyponym-hypernym relations are detected. Each hypernym is annotated with *isHypernymOf* to refer to its hyponyms within an IS schema. An example for such a relation is shown in Table I, where "worker" is a hypernym of "employee".

c) Homonyms: The term *homonym* is derived from the Greek word "homo", i.e., "equal", and describes words with the same syntax and pronunciation but different meaning. Thus, homonyms unite the cognates *homographs* (same syntax, different meaning) and *homophones* (same pronunciation, different meaning) [25]. Josko et al. [27] defined homonyms according to:

Definition 2 (Homonyms [27]): Let $sp : w(S) \times w(S) \rightarrow \{true, false\}$ be a function that returns if the graphy and pronunciation of two words within S are equal, according to LEX. Let $me : w(S) \times w(S) \rightarrow \{true, false\}$ be a function that returns if the meaning of two words within S are equal or nearly the same, according to LEX. A schema has homonyms *iff* $\exists w_i, w_j \in S$, where $i \neq j$, such that $sp(w_i, w_j) = true$ and $me(w_i, w_j) = false$. Homonyms are words that sound alike or are spelled alike, but have different meanings. The data defect "homonymous values" arises when homonymous terms are applied interchangeably and indicate the same fact about objects within a schema.

The majority of ISs in productive use implement the relational data model, and therefore lack a semantic annotation of the words within a schema. For example, the meaning of the word "bank", which can refer to the financial institution, or to a river bank, is not explicitly defined. The meaning is only implicitly available through the IS content, or known by domain experts. In such schemas, the distinction between homonyms and synonyms is not possible due to the lack of explicitly available semantics. Therefore, homonym detection is not part of our current implementation. However, more complex data models, like ontologies, would enable homonym detection. Part of our future work is to extend the readability metric with homonym detection.

C. A Metric to Measure Schema Readability

Based on the criteria from Section III-B, we suggest calculating the readability of an IS schema according to

$$Red(s) = \frac{\sum_{i=1}^{|w|} \#\text{fcrit}_i / \#\text{crit}}{|w|}, \tag{5}$$

where $|w|$ is the total number of words in schema s, $\#crit$ is the number of considered criteria, and $\#fcrit_i$ is the number of fulfilled criteria per word w_i. The metric delivers readability ratings that are normalized by [0,1], where 0.0 represents absolute poor readability, and 1.0 perfectly good readability. This characteristic aligns with the five requirements a sound DQ metric should fulfill by Heinrich et al. [28]. To discuss these requirements with respect to our readability metric, we calculated all possible ratings for a schema with 100 attributes. Figure 3 shows on the left side a boxplot, which indicates the distribution of the resulting metric ratings. On the right side of the figure, a line plot illustrates the metric rating per total number of fulfilled criteria, that is, number of attributes (100) multiplied by number of criteria (here 4: wordnet existence, case consistency, synonyms, and hypernyms).

The first requirement by [28] (*Existence of Minimum and Maximum Metric Values*) states that the metric results have to

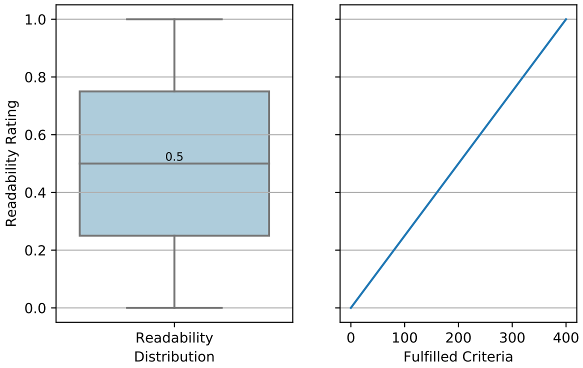

Fig. 3. Readability Metric Results

be normalized by [0,1], where 0.0 represents least readability, and 1.0 best readability. The whiskers of the boxplot in Figure 3 are bound by 0.0 and 1.0, which illustrates the fulfillment of this requirement. The second requirement (*Interval-Scaled Metric Values*) states that the steps of the metric result have to be equally spaced [28]. Both plots in Figure 3 show the fulfillment of this requirement, because (1) the median of all possible readability rating is with 0.5 the exact mean between the minimum and maximum values, and (2) for every fulfilled criteria per word, the gradient is increased with a fixed step size. If the readability rating of a schema is improved from 0.6 to 0.7, this corresponds to an improvement of the readability from 0.2 to 0.3. Consequently, the differences between the units of the metric results are always equally spaced. Although the fourth requirement (*Sound Aggregation of the Metric Values*) originally referred to the aggregation on IS data-level in terms of aggregating record-level QQ to table-level DQ, our metric also allows to aggregate the readability ratings between the single tables to an aggregated value for the entire IS schema. The remaining requirements R3 (*Quality of the Configuration Parameters and the Determination of the Metric Values*) and R5 (*Economic Efficiency of the Metric*) refer to the degree of automation, the parameters for the metric can be determined and measured with. We claim that both requirements are fulfilled, since we showed how to measure the criteria $crit$ in an automated way using WordNet.

IV. Proof-of-Concept Implementation

The readability metric proposed in this paper has been implemented and demonstrated in the Java-based DQ tool QualIe (Quality Assessment for Integrated Information Environments, pronounced [’kvɑlə]), introduced in [10]. QualIe automatically performs domain-independent quality measurement on both data-level and schema-level. Although the current version of the readability metric in QualIe was originally developed for the schema-level, it could be easily modified to assess the readability of string values on the content-level likewise. In this section, we demonstrate the functionality and applicability of our readability metric. For the interaction with WordNet, we used a Java WordNet API developed at the Massachusetts Institute of Technology (MIT) [29].

The selection of data sources for our proof-of-concept demonstration follows the evaluation suggestions for DQ metrics by Sadiq et al. [30], who promoted to use both, common synthetic data sets (for a manual verification of the readability calculation), as well as large real world data sets to show the applicability in practice. Thus, we selected the following DBs: (1) *Alphavantage* is highly volatile real-world stock exchange data, (2) *Chinook* [31] is a relational DB for digital media, (3) *Employees* [18] is a sample MySQL DB with six tables and about three million records that stores employees and departments within a company, (4) *Northwind* [32] is the well-known SQL DB from Microsoft, (5) *Metadynea* is a productive Cassandra DB from one of our industry partners that stores about 60 GB of chemometrics data distributed on three nodes, and (6) *Sakila* [33] is a MySQL sample DB for the administration of a film distribution with a more advanced schema (16 tables) than the employees DB. Table II, which is explained in the following sections, shows the readability ratings for each DB schema.

A. Alphavantage

We collected real-world stock exchange data with the alphavantage API [34], which yields a schema with information about the “time stamp”, “open” and “close” date, and the “volume” per stock. The observed table about IBM stock data achieves a quite high readability rating of 0.8750. Lowercases are used consistently in the entire schema. The main reason for the degraded readability is the attribute label “timestamp”, which has no exact match in WordNet, because the corresponding entry is “time stamp”. No synonyms or hypernyms are detected. However, additional cognates could be detected, if the attribute label “timestamp” would have been split into “time” and “stamp”, and both words are found in WordNet.

B. Chinook

The readability of the Chinook schema with 10 tables achieves the second-lowest rating with 0.5172. Lowercase is consistently used in the entire schema. A major point for the low readability are string concatenations. Several attribute labels have the table name as prefix, e.g., “customerid” or “artistid”. Here, an automated split during the preprocessing process is not possible, because no delimiter is used. Consequently, those attribute labels are treated as single words, which are not found in WordNet. The highest readability has the table `customer` (0.6731). It includes customer contact data, such as “email”, “phone”, and “address”, where the labels are single words that exist in the wordnet. The two synonyms “state” and “country” decrease the readability further.

C. Employees

The employees schema has a readability of 0.6902. The attribute labels are consistently written in lowercase and several labels are concatenated with an underscore, and therefore

TABLE II. READABILITY MEASUREMENTS

Schema	Readability	Concatenations	Cases	Abbreviations	Synonyms	Hypernyms
Alphavantage	0.8750	no split point	lower	-	-	-
Chinook	0.5172	no split point	lower	-	state ↔ country	-
Employees	0.6902	underscore	lower	-	-	first ← birth
Employees	0.8585	underscore	lower	file provided	-	first ← birth
Northwind	0.4247	no split point	lower	-	-	description ← picture; region ← country
Metadynea	0.9803	underscore	lower	-	-	level ← quality, intensity; time ← hour; type ← version
Sakila	0.9904	underscore	lower	-	duration ↔ length	code ← address; film ← feature

split during the preprocessing. In contrast to Chinook, this word concatenation does not lead to a deterioration of the readability, because both words are individually looked up in WordNet. For humans, the schema is easy readable due to the fact that most of the abbreviations are commonly used. For example, the abbreviations “dept” for departments and “emp” for employees are used as prefixes for attribute labels, such as “dept_name” and “emp_no”. However, the abbreviations “dept” and “emp” are not part of WordNet and therefore decrease the calculated readability. This issue can be resolved by including an abbreviations CSV file, which increases the readability ranking to 0.8585 (see Table II). An additional impact on the readability has the fact that “first” is recognized as hypernym of “birth”.

D. Northwind

The Northwind DB achieves with 0.4247 the lowest readability rating of all observed schemas, despite the fact that all words are consistently written in lowercase. Analogue to Chinook, the major reason for the low rating are string concatenations without delimiters. Many attribute labels include substrings, for example, “categoryname”, “companyname”, or “contacttitle”. Since no split point can be detected, those concatenations cannot be resolved.

E. Metadynea

The readability of the Metadynea schema is the second-best with 0.9803. All words are consistently written in lowercase and concatenated with underscores, which allows splitting. No unknown abbreviations are used and all words are included in WordNet. The only drawbacks found are several hypernyms, e.g., “time” is a hypernym of “hour”.

F. Sakila

The overall best readability rating with 0.9904 is achieved by Sakila, where all words except “username” are included in WordNet. Labels used for attributes are consistently written in lowercase. One minor problem is the attribute label “address2”, which can neither be splitted nor has a match in WordNet. Further, several cognates are detected, e.g., the word “code” contained in the attribute label “postal code” is identified as hypernym of the word “address”.

V. CONCLUSION

The readability of IS schemas is of particular importance to ensure automated schema integration and to allow humans a correct interpretation of table and attribute names. In this paper, we have introduced a novel metric for the readability of IS schemas, which is based on a set of readability critera that are applied to words extracted from a schema. In the current state, the metric considers the criteria (a) entry in a wordnet, (b) consistency of cases, and the cognates (c) synonyms and (d) hypernyms. To demonstrate the applicability of our metric, we implemented it in the DQ tool QuaIIe and measured the readability of multiple synthetic and real data sources.

In our ongoing and future work, we plan to extend the readability metric with (1) text-based approaches, (2) string similarity, (3) normalization with respect to the schema size, as well as (4) further investigation on string splitting. Since there is a lot of related work about readability concerning text-based approaches, it could also be beneficial to take into account word complexity [13]. Words with many syllables are more complex and a schema with a lot of complex words is less readable. The second possible improvement with string similarity could be used to detect similar attribute names that are only distinguished by a typo, for example, “productNumber” and “porductNumber”. Those words would not be considered similar with the presented algorithm, but could be taken into account with string similarity algorithms, like the Levensthein distance. The current implementation does not consider the size of the evaluated IS. Since larger ISs tend to have more (readability) errors, an interesting index could be the consideration of errors per hundred tables. In addition, we think that the challenging topic of splitting strings without a clear split point, e.g., “categoryname”, would be worth to be investigated in the future. Last, but not least, we are going to extend and refine the evaluation of our metric by (1) additionally using benchmark data sets for federated ISs, and (2) conducting a user study to compare the readability ratings of the metric to the assessment of real users.

ACKNOWLEDGMENT

The research reported in this paper has been partly supported by the Austrian Ministry for Transport, Innovation and Technology, the Federal Ministry of Digital and Economic Affairs, and the Province of Upper Austria in the frame of the COMET center SCCH.

REFERENCES

[1] W. W. Eckerson, “Data Quality and the Bottom Line – Achieving Business Success through a Commitment to High Quality Data,” The Data Warehousing Institute, Technical Report, 2002.

[2] D. Loshin, *The Practitioners Guide to Data Quality Improvement*. Elsevier Inc., 2011.

[3] Y. Wand and R. Y. Wang, “Anchoring Data Quality Dimensions in Ontological Foundations,” *Communications of the ACM*, vol. 39, no. 11, Nov. 1996, pp. 86–95.

[4] C. Batini and M. Scannapieco, *Data and Information Quality: Concepts, Methodologies and Techniques*. Springer International Publishing, 2016.

[5] L. Cai and Y. Zhu, “The Challenges of Data Quality and Data Quality Assessment in the Big Data Era,” in *Data Science Journal*, vol. 14. Ubiquity Press, 2015, pp. 1–10.

[6] N. A. Emran, S. Embury, P. Missier, M. N. M. Isa, and A. K. Muda, “Measuring Data Completeness for Microbial Genomics Database,” in *5th Asian Conference on Intelligent Information and Database Systems*. Springer-Verlag Berlin Heidelberg, 2013, pp. 186–195.

[7] O. Foley and M. Helfert, “The Development of an Objective Metric for the Accessibility Dimension of Data Quality,” in *4th International Conference on Innovations in Information Technology*. IEEE, 2007, pp. 11–15.

[8] L. Pipino, Y. Lee, and R. Y. Wang, “Data Quality Assessment,” in *Communications Of The ACM*. ACM New York, April 2002, pp. 211–218.

[9] G. Vossen, *Datenmodelle, Datenbanksprachen und Datenbankmanagementsysteme [Data Models, Database Languages, and Database Management Systems]*. Oldenbourg Verlag, 2008.

[10] L. Ehrlinger, B. Werth, and W. Wöß, “QuaIIe: A Data Quality Assessment Tool for Integrated Information Systems,” *Proceedings of the Tenth International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA 2018)*, 2018, pp. 21–31.

[11] R. Y. Wang and D. M. Strong, “Beyond Accuracy: What Data Quality Means to Data Consumers,” *Journal of Management Information Systems*, vol. 12, no. 4, March 1996, pp. 5–33.

[12] J. Zhao and M. Kan, “Domain-Specific Iterative Readability Computation,” in *Proceedings of the 10th annual joint conference on Digital libraries*. ACM New York, June 2010, pp. 205–214.

[13] R. J. Senter and E. A. Smith, “Automated Readability Index,” United States Air Force Aerospace Medical Research Laboratories, Technical Report, *AMRLTR-66-220*, November 1967.

[14] A. D. Renzisa *et al.*, “A Domain Independent Readability Metric for Web Service Descriptions,” in *Computer Standards and Interfaces*. Elsevier, 2017, pp. 124–141.

[15] S. Hoberman, *Data Model Scorecard*. Technics Publications, 2015.

[16] X. Yan, D. Song, and X. Li, “Concept-based Document Readability in Domain Specific Information Retrieval,” in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM ’06)*. NY, USA: ACM, 2006, pp. 540–549.

[17] L. Ehrlinger and W. Wöß, “Semi-Automatically Generated Hybrid Ontologies for Information Integration,” in *Joint Proceedings of the Posters and Demos Track of 11th International Conference on Semantic Systems*. CEUR Workshop Proceedings, 2015, pp. 100–104.

[18] Oracle Corporation, “Employees Sample Database,” https://dev.mysql.com/doc/employee/en [retrieved: April, 2019].

[19] The Global WordNet Association, “Global Wordnet Association,” http://globalwordnet.org/ [retrieved: April, 2019].

[20] DBpedia Association, “DBpedia,” http://wiki.dbpedia.org [retrieved: April, 2019], 2018.

[21] “WordNet du Franais,” https://wonef.fr [retrieved: April, 2019].

[22] Princeton University, “WordNet - A Lexical Database for English,” https://wordnet.princeton.edu [retrieved: April, 2019].

[23] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, “Introduction to WordNet: An On-line Lexical Database,” in *International Journal of Lexicography*, vol. 3. Oxford University Press, 12 1990, pp. 235 – 244.

[24] M. A. Finlayson, “Java Libraries for Accessing the Princeton Wordnet: Comparison and Evaluation,” in *Proceedings of the 7th International Global WordNet Conference*, H. Orav, C. Fellbaum, and P. Vossen, Eds. Association for Computational Linguistics, January 2014, pp. 78–85.

[25] Oxford University Press, “Oxford Dictionaries,” https://en.oxforddictionaries.com/definition [retrieved: April, 2019].

[26] L. Ehrlinger, B. Werth, and W. Wöß, “Automated Continuous Data Quality Measurement with QuaIIe,” *International Journal on Advances in Software*, vol. 11, no. 3 & 4, 2018, pp. 400–417.

[27] J. M. B. Josko, M. K. Oikawa, and J. E. Ferreira, “A Formal Taxonomy to Improve Data Defect Description,” in *Database Systems for Advanced Applications*, H. Gao, J. Kim, and Y. Sakurai, Eds. Springer International Publishing, 2016, pp. 307–320.

[28] B. Heinrich, D. Hristova, M. Klier, A. Schiller, and M. Szubartowicz, “Requirements for Data Quality Metrics,” *Journal of Data and Information Quality*, vol. 9, no. 2, January 2018, pp. 12:1–12:32.

[29] Massachusetts Institute of Technology, “The MIT Java Wordnet Interface,” https://projects.csail.mit.edu/jwi [retrieved: April, 2019].

[30] S. Sadiq *et al.*, “Data Quality: The Role of Empiricism,” *ACM SIGMOD Record*, vol. 46, no. 4, 2018, pp. 35–43.

[31] Microsoft Inc., “ChinookDatabase,” https://archive.codeplex.com/?p=chinookdatabase [retrieved: April, 2019].

[32] Microsoft Inc., “Northwind and pubs Sample Databases for SQL Server 2000,” 2018, https://www.microsoft.com/en-us/download/details.aspx?id=23654 [retrieved: April, 2019].

[33] Oracle Corporation, “Sakila Sample Database,” https://dev.mysql.com/doc/sakila/en [retrieved: April, 2019].

[34] Alpha Vantage Inc., “ALPHA VANTAGE,” 2018, https://www.alphavantage.co [retrieved: April, 2019].

ISBN: 978-1-61208-715-3