An Evaluation of Metadata and Data Quality

on Person-Level, Aggregated, Thesauri, Statistical Classifications, and Rectangular Data Sets

Thomas Bosch[1], Benjamin Zapilko[1], Joachim Wackerow[1], and Kai Eckert[2]

[1] GESIS – Leibniz Institute for the Social Sciences, Germany
{firstname.lastname}@gesis.org,
[2] University of Mannheim, Germany
kai@informatik.uni-mannheim.de

Abstract. From 2012 to 2015 together with other Linked Data community members and experts from the social, behavioural, and economic sciences (*SBE*), we developed diverse vocabularies to represent *SBE* metadata and rectangular data in RDF. The *DDI-RDF Discovery Vocabulary (Disco)* is designed to support the dissemination, management, and reuse of person-level data, i.e., data about individuals, households, and businesses, collected in form of responses to studies and archived for research purposes. The *RDF Data Cube Vocabulary (Data Cube)* is a W3C recommendation for expressing *data cubes*, i.e. multi-dimensional aggregate data. *Physical Data Description (PHDD)* is a vocabulary to model data in rectangular format. The data could either be represented in records with character-separated values (*CSV*) or fixed length. The *Simple Knowledge Organization System (SKOS)* is a vocabulary to build knowledge organization systems such as thesauri, classification schemes, and taxonomies. *XKOS* is a SKOS extension to describe formal statistical classifications.

To ensure high quality of and trust in both metadata and data, their representation in RDF must satisfy certain criteria - specified in terms of RDF constraints. In this paper, we evaluated the metadata and data quality of large real world aggregated (*QB*), person-level (*Disco*), thesauri (*SKOS*), rectangular (*PHDD*), and statistical classification (*XKOS*) data sets by means of RDF constraints. RDF Constraints are instances of RDF constraint types either corresponding to RDF validation requirements or to data model specific constraint types. We validated more than 4.2 billion triples and 15 thousand data sets using the *RDF Validator*, a validation environment which is available at http://purl.org/net/rdfval-demo.

Keywords: RDF Validation, RDF Constraints, DDI-RDF Discovery Vocabulary, RDF Data Cube Vocabulary, Thesauri, SKOS, Rectangular Data, Statistical Classifications, XKOS, Linked Data, Semantic Web

1 RDF Validation of Metadata and Data

Bosch et al. identified in total 74 requirements to formulate RDF constraints; each of them corresponding to a constraint type. We published a technical report[3] in which we explain each requirement (constraint type) in detail and give examples for each (represented by different constraint languages). The knowledge representation formalism *Description logics (DL)*, with its well-studied theoretical properties, provides the foundational basis for each constraint type. Therefore, this technical report contains mappings to DL to logically underpin each requirement and to determine which DL constructs are needed to express each constraint type [2]. We recently published a technical report in which we describe constraints to validate metadata on person-level, aggregated data, and thesauri. We assign each constraint to constraint types corresponding to RDF validation requirements or to data model specific constraint types[4] [3].

We distinguish two validation types: (1) *Content-Driven Validation* $\mathcal{C}_C$ contains the set of constraints ensuring that the data is consistent with the intended syntax, semantics, and integrity of given data models. (2) *Technology-Driven Validation* $\mathcal{C}_T$ includes the set of constraints which can be generated automatically out of data models, such as cardinality restrictions, universal and existential quantifications, domains, and ranges. We determined the default *severity level* (corresponds to requirement *R-158*) for each constraint to indicate how serious the violation of the constraint is. We propose an extensible metric to measure the continuum of severity levels ranging from $\mathcal{SL}_0$ (informational) via $\mathcal{SL}_1$ (warning) to $\mathcal{SL}_2$ (error). Although we provide default severity levels for each constraint, users should be able to specify severity levels of constraints they need to validate for their individual use cases, i.e., users should be able to define use case specific severity levels for constraints.

2 Evaluation

We exhaustively evaluated the metadata quality of large real world aggregated (*QB*), person-level (*Disco*), and thesauri (*SKOS*) data sets by means of both $\mathcal{C}_C$ and $\mathcal{C}_T$ constraints of the majority of the constraint types. We validated 9,990 / 3,775,983,610 (*QB*), 4,178 / 477,737,281 (*SKOS*), and 1,526 / 9,673,055 (*Disco*) data sets / triples using the *RDF Validator*[5] (available at http://purl.org/net/rdfval-demo) in batch mode. That are more than 4.2 billion triples and 15 thousand data sets. We validated, i.a., (1) *QB* data sets published by the *Australian Bureau of Statistics (ABS)*, the *European Central Bank (ECB)*, and the *Organisation for Economic Co-operation and Development (OECD)*, (2) *SKOS* thesauri like the *AGROVOC Multilingual agricultural thesaurus*, the *STW Thesaurus for Economics*, and the *Thesaurus for the Social Sciences (TheSoz)*, and

[3] Available at: http://arxiv.org/abs/1501.03933

[4] Requirements/Constraint types and constraints are uniquely identified by alphanumeric technical identifiers like *R-1*

[5] For details about the *RDF Validator* see [1]

(3) *Disco* data sets provided by the *Microdata Information System (Missy)*, the *DwB Discovery Portal*, the *Danish Data Archive (DDA)*, and the *Swedish National Data Service (SND)*. As we evaluated nearly 10 thousand *QB* data sets, we published the evaluation results for each data set in form of one document per SPARQL endpoint[6]. The correctness of all constraints, i.e., the gold standard, has been proved by SBE domain experts. Table 1 shows the evaluation results.

Criteria[7]	***Disco***	***QB***	***SKOS***	**Total**
Triples	9,673,055	3,775,983,610	477,737,281	4,263,393,946
Data Sets	1,526	9,990	4,178	15,694
CV	3,545,703	45,635,846	5,540,988	54,722,537
CV ($\mathcal{SL}_0$)	2,437,922 (68.8%)	0 (0%)	2,281,740 (41.2%)	4,719,662 (8.6%)
CV ($\mathcal{SL}_1$)	473,574 (13.4%)	45,520,613 (99.75%)	3,259,248 (58.8%)	49,253,435 (90%)
CV ($\mathcal{SL}_2$)	634,207 (17.9%)	115,233 (0.25%)	0 (0%)	749,440 (1.4%)
CT	52 (15|37)[8]	20 (7|13)	14 (4|10)	53
CT ($\mathcal{C}_C$)	30 (57.7%)	5 (25%)	5 (35.7%)	30 (56.6%)
CT ($\mathcal{C}_T$)	22 (42.3%)	15 (75%)	9 (64.3%)	23 (43.4%)
C	142 (77|65)	35 (20|15)	35 (17|18)	212
C ($\mathcal{C}_C$)	72 (50.7%)	16 (45.7%|12|4)	21 (60%|13|8)	109 (51.4%)
C ($\mathcal{C}_T$)	70 (49.3%)	19 (54.3%|8|11)	14 (40%|4|10)	103 (48.6%)
C ($\mathcal{SL}_0$)	75 (52.8%|44|31)	4 (11.4%|0|4)	21 (60%|12|9)	100 (47.2%)
C ($\mathcal{SL}_1$)	9 (6.3%|8|1)	3 (8.6%|3|0)	5 (14.3%|5|0)	17 (8%)
C ($\mathcal{SL}_2$)	58 (40.8%|25|33)	28 (80%|17|11)	9 (25.7%|0|9)	95 (44.8%)

Table 1: Evaluation

We identified 142 *Disco* constraints ($\mathcal{C}_C$ and $\mathcal{C}_T$ constraints to the same extend) assigned to 52 distinct constraint types and implemented 77 of them to actually validate person-level data sets. For *QB*, we specified more $\mathcal{C}_T$ (54%) than $\mathcal{C}_C$ constraints; for *SKOS*, however, more $\mathcal{C}_C$ constraints (60%). We instantiated more $\mathcal{C}_C$ (58%) than $\mathcal{C}_T$ constraint types to define *Disco* constraints; for *QB* (75%) and *SKOS* (64%), on the other side, more $\mathcal{C}_T$ constraint types. In total, we used 53 of overall 82 distinct constraint types (57% of them are $\mathcal{C}_C$ constraint types) to define 212 constraints (equally $\mathcal{C}_C$ and $\mathcal{C}_T$ constraints).

For *Disco* and *SKOS*, more than the half of the constraints are associated with the weakest severity level $\mathcal{SL}_0$. Within the context of *QB*, 80% of the constraints are classified as the most serious ones ($\mathcal{SL}_2$). All in all, there are a little bit more $\mathcal{SL}_0$ then $\mathcal{SL}_2$ constraints, whereas $\mathcal{SL}_1$ constraints are negligible. *Existential quantifications* (32.4%, *Disco*), *data model consistency* (31.4%, *QB*), and *structure* (28.6%, *SKOS*) are the constraint types the most constraints are instantiated from. By validating *QB* data sets, we got the most

[6] Available at: https://github.com/boschthomas/rdf-validation/tree/master/evaluation/data-sets/data-cube

[7] *C (constraints), CT (constraint types), CV (constraint violations)*

[8] (implemented |not yet implemented)

constraint violations (more than 45 millions), followed by *SKOS* and *Disco* (with more than 5.5 and 3.5 millions) - consequently, almost 55 million constraint violations were raised during the evaluation which could be used to enhance the metadata quality of these data sets. Close to 70% of all *Disco* constraint violations are caused by violating $\mathcal{SL}_0$ constraints. For *QB* (nearly 100%) and *SKOS* (almost 60%), the majority of the raised constraint violations are classified to be more serious ($\mathcal{SL}_1$). 80% of all *QB* constraints are $\mathcal{SL}_2$ constraints leading to less than 1% of all *QB* constraint violations. Altogether, exactly 90% of the constraint violations are assigned to the severity level $\mathcal{SL}_1$. These findings are surprising as only 8% of all defined constraints are $\mathcal{SL}_1$ constraints. The constraints responsible for the largest numbers of constraint violations are *DISCO-C-LABELING-AND-DOCUMENTATION-06* and *DISCO-C-COMPARISON-VARIABLES-02* (both 547,916) (*Disco*), *DATA-CUBE-C-DATA-MODEL-CONSISTENCY-05* (45,514,102) (*QB*), and *SKOS-C-LANGUAGE-TAG-CARDINALITY-01* (2,508,903) (*SKOS*).

2.1 Legend

In this section, we describe how the tables in this paper should be read. Table 2 gives an overview over the symbols used in subsequent tables.

Symbol	Description
✓	Validation Successful (without any constraint violation)
X	Constraint Violations
>X	Poor Performance/Scaling
✗	Very Poor Performance/Scaling
(!)	Not Yet Implemented Constraint
(X)	The validation of *X* data sets could not be finished, due to SPARQL endpoints' technical restrictions (e.g. defined timeouts).
*	default severity level $\mathcal{SL}_0$ (informational)
**	default severity level $\mathcal{SL}_1$ (warning)
***	default severity level $\mathcal{SL}_2$ (error)

Table 2: Legend

- **Constraint Violations.** When constraints are violated, *X* indicates the number of raised constraint violation triples.
- **Poor Performance/Scaling.** The performance of the implementation of the underlying SPARQL CONSTRUCT query is too poor to get all resulting constraint violation triples. Therefore, a limit of *X* result constraint violation triples is set. It is likely that there are more than *X* constraint violations. Although, the result set contains not the whole set of raised constraint violation triples, the constraint can be used as an indicator if there is data

not conforming to the constraint and to resolve constraint violations step by step. As part of future work, the performance will be improved.
– **Very Poor Performance/Scaling.** The performance of the implementation of the underlying SPARQL CONSTRUCT query is too poor to get any results, even though a limit of result constraint violation triples is set. As part of future work, the performance will be improved.

3 Evaluation of Person-Level Metadata (*Disco*)

In this section, the quality of the metadata on person-level (*Disco*) data sets is evaluated by validating appropriate RDF constraints assigned to several RDF constraint types. First, we show the results of the evaluation of diverse data sets, then we give an overview over the evaluated data sets, and finally we provide details about the evaluation.

3.1 Evaluation Results

Table 3 shows the results of the evaluation of *Disco* data sets.

Evaluation Criteria	**Counts**
Validated Triples	9,673,055
Validated Data Sets	1,526
Constraint Violations	3,545,703
Constraint Violations ($\mathcal{SL}_0$)	2,437,922 (68.8%)
Constraint Violations ($\mathcal{SL}_1$)	473,574 (13.4%)
Constraint Violations ($\mathcal{SL}_2$)	634,207 (17.9%)
Constraint (Most Constraint Violations)	*DISCO-C-LABELING-AND-DOCUMENTATION-06* (547,916)
	DISCO-C-COMPARISON-VARIABLES-02 (547,916)
Constraint (Most Constraint Violations ($\mathcal{SL}_0$))	*DISCO-C-LABELING-AND-DOCUMENTATION-06* (547,916)
Constraint (Most Constraint Violations ($\mathcal{SL}_1$))	*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-46* (468,807)
Constraint (Most Constraint Violations ($\mathcal{SL}_2$))	*DISCO-C-COMPARISON-VARIABLES-02* (547,916)
Constraint Types	52 (15|37)[9]
Constraint Types ($\mathcal{C}_C$)	30 (57.7%)
Constraint Types ($\mathcal{C}_T$)	22 (42.3%)
Constraint Types (Most Constraints)	1. Existential Quantifications: 46 (32.4%|46|0)[10]
	2. Data Model Consistency: 7 (1|6)
	3. Aggregation: 7 (0|7)
Constraint Type (Most Constraints ($\mathcal{SL}_2$))	Existential Quantifications: 9 (9|0)
Constraints	142 (77|65)
Constraints ($\mathcal{C}_C$)	72 (50.7%)
Constraints ($\mathcal{C}_T$)	70 (49.3%)
Constraints ($\mathcal{SL}_0$)	75 (52.8%|44|31)
Constraints ($\mathcal{SL}_1$)	9 (6.3%|8|1)
Constraints ($\mathcal{SL}_2$)	58 (40.8%|25|33)

Table 3: Evaluation of Disco Data Sets - Evaluation Results

3.2 Data Sets Overview

Tables 4 and 6 give an overview over the evaluated *Disco* data sets, their abbreviations, and publicly available SPARQL endpoints. Table 5 comprehends the number of triples, data sets, and instances of multiple vocabulary-specific classes.

Abbr.	**Disco Data Sets**
Missy	*Microdata Information System*[11]
DwB	*DwB Discovery Portal*[12]
DDA-SND	*DDI-RDF*[13]
	provided by the *Danish Data Archive (DDA)*[14] and Swedish National Data Service (SND)[15]

Table 4: Disco Data Sets Abbreviations

[9] legend: absolute number (absolute number implemented |absolute number not yet implemented)
[10] legend: absolute number (percentage value |absolute number implemented |absolute number not yet implemented)
[11] http://www.gesis.org/missy/eu/missy-home
[12] http://dwb-dev.nsd.uib.no/portal
[13] http://ddi-rdf.borsna.se/
[14] http://samfund.dda.dk/dda/default-en.asp
[15] http://snd.gu.se/en

	Counts									
Data Sets	**triples**	**disco:StudyGroup**	**disco:Study**	**disco:LogicalDataSet**	**disco:Universe**	**disco:Variable**	**disco:Question**	**disco:SummaryStatistics**	**disco:CategoryStatistics**	**skos:Concept**
Missy	5,068,838	6	45	159	1,125	21,040	0	0	0	147,193
DwB	2,332,802	0	1,387	1,367	2,796	446,806	0	0	0	0
DDA-SND	2,271,415	0	1,490	0	10,188	80,070	139,237	0	0	290,963
Total	9,673,055			1,526						

Table 5: Disco Data Sets Overview

Data Sets	**SPARQL Endpoint**
Missy	http://svko-missy:8181/openrdf-workbench/repositories/native-java-store/summary
DwB	http://dwb-dev.nsd.uib.no/sparql
DDA-SND	http://ddi-rdf.borsna.se/endpoint/

Table 6: Disco SPARQL Endpoints

3.3 Detailed Evaluation

In this sub section, we give details about the evaluation in form of diverse tables containing the number of constraint violations per evaluated data set and constraint of particular constraint types.

	Data Sets		
Existential Quantifications (1)	*Missy*	*DwB*	*DDA-SND*
DISCO-C-EXISTENTIAL-QUANTIFICATIONS-01***	✓	✓	✓
DISCO-C-EXISTENTIAL-QUANTIFICATIONS-02***	7	17	1,490
DISCO-C-EXISTENTIAL-QUANTIFICATIONS-03*	✓	✓	✓
DISCO-C-EXISTENTIAL-QUANTIFICATIONS-04*	11,021	445,381	62,260
DISCO-C-EXISTENTIAL-QUANTIFICATIONS-05*	✓	✓	139,237
DISCO-C-EXISTENTIAL-QUANTIFICATIONS-06*	12	1,367	✓
DISCO-C-EXISTENTIAL-QUANTIFICATIONS-07*	6	✓	✓
DISCO-C-EXISTENTIAL-QUANTIFICATIONS-08*	45	1,387	1,490
DISCO-C-EXISTENTIAL-QUANTIFICATIONS-09*	6	✓	✓
DISCO-C-EXISTENTIAL-QUANTIFICATIONS-10*	45	1,387	1,490

Table 7: Evaluation of Disco Data Sets - Existential Quantifications (1)

	Data Sets		
Existential Quantifications (2)	*Missy*	*DwB*	*DDA-SND*
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-11**	6	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-12**	6	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-13**	✓	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-14**	45	1,387	1,490
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-15**	45	1,387	1,490
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-16**	✓	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-17**	159	1,367	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-18**	159	1,367	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-19**	✓	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-20**	✓	1,367	✓

Table 8: Evaluation of Disco Data Sets - Existential Quantifications (2)

	Data Sets		
Existential Quantifications (3)	*Missy*	*DwB*	*DDA-SND*
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-21**	✓	1,367	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-22**	✓	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-23**	6	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-24**	45	1,387	1,490
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-25**	45	1,387	1,490
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-26**	45	1,387	1,490
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-27****	✓	130	1,490
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-28***	159	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-29***	✓	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-30***	✓	✓	✓

Table 9: Evaluation of Disco Data Sets - Existential Quantifications (3)

Existential Quantifications (4)	Data Sets		
	Missy	*DwB*	*DDA-SND*
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-31***	159	1,367	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-32****	✓	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-33****	✓	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-34****	✓	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-35****	✓	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-36****	✓	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-37**	18,625	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-38**	✓	✓	750
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-39****	✓	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-40**	✓	✓	139,237

Table 10: Evaluation of Disco Data Sets - Existential Quantifications (4)

Existential Quantifications (5)	Data Sets		
	Missy	*DwB*	*DDA-SND*
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-41**	✓	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-42**	✓	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-43**	15,733	446,806	80,070
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-44**	159	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-45**	6,784	446,806	19,221
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-46***	11,550	446,806	10,451

Table 11: Evaluation of Disco Data Sets - Existential Quantifications (5)

	Data Sets		
Conditional Properties	*Missy*	*DwB*	*DDA-SND*
DISCO-C-CONDITIONAL-PROPERTIES-01***	✓	✓	80,070
DISCO-C-CONDITIONAL-PROPERTIES-02**	12	✓	✓
DISCO-C-CONDITIONAL-PROPERTIES-03**	90	✓	2,980
DISCO-C-CONDITIONAL-PROPERTIES-04***	6	✓	✓
DISCO-C-CONDITIONAL-PROPERTIES-05***	45	1,387	1,490
DISCO-C-CONDITIONAL-PROPERTIES-06***	✓	✓	✓

Table 12: Evaluation of Disco Data Sets - Conditional Properties

	Data Sets		
Provenance	*Missy*	*DwB*	*DDA-SND*
DISCO-C-PROVENANCE-01*	6	✓	✓
DISCO-C-PROVENANCE-02*	45	1,387	1,490
DISCO-C-PROVENANCE-03*	159	1,367	✓
DISCO-C-PROVENANCE-04*	✓	1,367	✓

Table 13: Evaluation of Disco Data Sets - Provenance

	Data Sets		
Labeling and Documentation	*Missy*	*DwB*	*DDA-SND*
DISCO-C-LABELING-AND-DOCUMENTATION-01[*]	6	✓	✓
DISCO-C-LABELING-AND-DOCUMENTATION-02[*]	45	1,387	1,490
DISCO-C-LABELING-AND-DOCUMENTATION-03[*]	159	1,367	✓
DISCO-C-LABELING-AND-DOCUMENTATION-04[*]	✓	1,367	✓
DISCO-C-LABELING-AND-DOCUMENTATION-05[*]	✓	✓	✓
DISCO-C-LABELING-AND-DOCUMENTATION-06[*]	21,040	446,806	80,070

Table 14: Evaluation of Disco Data Sets - Labeling and Documentation

	Data Sets		
Data Model Consistency	*Missy*	*DwB*	*DDA-SND*
DISCO-C-DATA-MODEL-CONSISTENCY-01 (!)[***]			
DISCO-C-DATA-MODEL-CONSISTENCY-02 (!)[***]			
DISCO-C-DATA-MODEL-CONSISTENCY-03 (!)[***]			
DISCO-C-DATA-MODEL-CONSISTENCY-04 (!)[***]			
DISCO-C-DATA-MODEL-CONSISTENCY-05[***]	✓	✓	✓
DISCO-C-DATA-MODEL-CONSISTENCY-06 (!)[***]			
DISCO-C-DATA-MODEL-CONSISTENCY-07 (!)[***]			

Table 15: Evaluation of Disco Data Sets - Data Model Consistency

	Data Sets		
Comparison	*Missy*	*DwB*	*DDA-SND*
*DISCO-C-COMPARISON-VARIABLES-01 (!)***			
*DISCO-C-COMPARISON-VARIABLES-02****	21,040	446,806	80,070
*DISCO-C-COMPARISON-VARIABLES-03 (!)****			
*DISCO-C-COMPARISON-VARIABLES-04**	18,625	✓	✓
*DISCO-C-COMPARISON-VARIABLES-05****	159	✓	✓

Table 16: Evaluation of Disco Data Sets - Comparison

	Data Sets		
Mathematical Operations	*Missy* *DwB*		*DDA-SND*
*DISCO-C-MATHEMATICAL-OPERATIONS-01 (!)****			
*DISCO-C-MATHEMATICAL-OPERATIONS-02 (!)****			
*DISCO-C-MATHEMATICAL-OPERATIONS-03 (!)****			
*DISCO-C-MATHEMATICAL-OPERATIONS-04 (!)****			
*DISCO-C-MATHEMATICAL-OPERATIONS-05 (!)****			

Table 17: Evaluation of Disco Data Sets - Mathematical Operations

	Data Sets	
Language Tags	*Missy DwB*	*DDA-SND*
*DISCO-C-LANGUAGE-TAG-MATCHING-01 (!)**		
*DISCO-C-LANGUAGE-TAG-CARDINALITY-01 (!)**		
*DISCO-C-LANGUAGE-TAG-CARDINALITY-02 (!)**		
*DISCO-C-LANGUAGE-TAG-CARDINALITY-03 (!)**		

Table 18: Evaluation of Disco Data Sets - Language Tags

	Data Sets	
Aggregation	*Missy DwB*	*DDA-SND*
*DISCO-C-AGGREGATION-01 (!)**		
*DISCO-C-AGGREGATION-02 (!)**		
*DISCO-C-AGGREGATION-03 (!)**		
*DISCO-C-AGGREGATION-04 (!)**		
*DISCO-C-AGGREGATION-05 (!)**		
*DISCO-C-AGGREGATION-06 (!)**		
*DISCO-C-AGGREGATION-07 (!)**		

Table 19: Evaluation of Disco Data Sets - Aggregation

	Data Sets		
Disco Constraints	*Missy*	*DwB*	*DDA-SND*
DISCO-C-ALLOWED-VALUES-01***	✓	✓	✓
DISCO-C-LITERAL-RANGES-01***	✓	✓	✓
DISCO-C-INVERSE-FUNCTIONAL-PROPERTIES-01***	✓	✓	✓
DISCO-C-INVERSE-FUNCTIONAL-PROPERTIES-02***	✓	✓	✓
DISCO-C-CLASS-SPECIFIC-PROPERTY-RANGE-01***	✓	✓	✓
DISCO-C-MEMBERSHIP-IN-CONTROLLED-VOCABULARIES-01***	✓	✓	✗
DISCO-C-LITERAL-VALUE-COMPARISON-01***	✓	1,299	✓
DISCO-C-CONTEXT-SPECIFIC-VALID-PROPERTIES-01*	21,038	✓	✓
DISCO-C-DATA-PROPERTY-FACETS-01**	✓	✓	✓
DISCO-C-DATA-PROPERTY-FACETS-02**	✓	✓	✓

Table 20: Evaluation of Disco Data Sets - Disco Constraints (1)

	Data Sets		
Disco Constraints	*Missy*	*DwB*	*DDA-SND*
DISCO-C-VALUE-IS-VALID-FOR-DATATYPE-01***	30	6,932	✓
DISCO-C-VALUE-IS-VALID-FOR-DATATYPE-02***	✓	✓	✓
DISCO-C-SUBSUMPTION-01 (!)$^{***\text{B}}$			
DISCO-C-CLASS-EQUIVALENCE-01 (!)*			
DISCO-C-SUB-PROPERTIES-01 (!)***			
DISCO-C-PROPERTY-DOMAIN-01 (!)***			
DISCO-C-PROPERTY-RANGES-01 (!)***			
DISCO-C-INVERSE-OBJECT-PROPERTIES-01 (!)***			
DISCO-C-INVERSE-OBJECT-PROPERTIES-02 (!)***			
DISCO-C-INVERSE-OBJECT-PROPERTIES-03 (!)***			
DISCO-C-DISJOINT-PROPERTIES-01 (!)***			

Table 21: Evaluation of Disco Data Sets - Disco Constraints (2)

	Data Sets	
Disco Constraints	*Missy DwB*	*DDA-SND*
DISCO-C-ASYMMETRIC-OBJECT-PROPERTIES-01 (!)***		
DISCO-C-IRREFLEXIVE-OBJECT-PROPERTIES-01 (!)***		
DISCO-C-CLASS-SPECIFIC-IRREFLEXIVE-OBJECT-PROPERTIES-01 (!)***		
DISCO-C-CLASS-SPECIFIC-IRREFLEXIVE-OBJECT-PROPERTIES-02 (!)***		
DISCO-C-DISJOINT-CLASSES-01 (!)***		
DISCO-C-EQUIVALENT-PROPERTIES-01 (!)*		
DISCO-C-LITERAL-PATTERN-MATCHING-01 (!)*		
DISCO-C-DISJUNCTION-01 (!)***		
DISCO-C-UNIVERSAL-QUANTIFICATIONS-01 (!)***		
DISCO-C-MINIMUM-QUALIFIED-CARDINALITY-RESTRICTIONS-01 (!)***		

Table 22: Evaluation of Disco Data Sets - Disco Constraints (3)

Disco Constraints	Data Sets: *Missy DwB*	Data Sets: *DDA-SND*
DISCO-C-MAXIMUM-QUALIFIED-CARDINALITY-RESTRICTIONS-01 (!)***		
DISCO-C-EXACT-QUALIFIED-CARDINALITY-RESTRICTIONS-01 (!)***		
DISCO-C-CONTEXT-SPECIFIC-EXCLUSIVE-OR-OF-PROPERTY-GROUPS-01 (!)*		
DISCO-C-IRI-PATTERN-MATCHING-01 (!)*		
DISCO-C-ORDERING-01 (!)*		
DISCO-C-ORDERING-02 (!)*		
DISCO-C-ORDERING-03 (!)*		
DISCO-C-STRING-OPERATIONS-01 (!)*		
DISCO-C-CONTEXT-SPECIFIC-VALID-CLASSES-01 (!)*		
DISCO-C-CONTEXT-SPECIFIC-VALID-PROPERTIES-01 (!)*		

Table 23: Evaluation of Disco Data Sets - Disco Constraints (4)

Disco Constraints	Data Sets: *Missy DwB*	Data Sets: *DDA-SND*
DISCO-C-DEFAULT-VALUES-01 (!)[*]		
DISCO-C-WHITESPACE-HANDLING-01 (!)[*]		
DISCO-C-HTML-HANDLING-01 (!)[*]		
DISCO-C-HTML-HANDLING-02 (!)[*]		
DISCO-C-RECOMMENDED-PROPERTIES-01 (!)[*]		
DISCO-C-HANDLE-RDF-COLLECTIONS-01 (!)[*]		
DISCO-C-HANDLE-RDF-COLLECTIONS-02 (!)[*]		
DISCO-C-USE-SUB-SUPER-RELATIONS-IN-VALIDATION-01 (!)[*]		
DISCO-C-USE-SUB-SUPER-RELATIONS-IN-VALIDATION-02 (!)[*]		
DISCO-C-STRUCTURE-01 (!)[***]		

Table 24: Evaluation of Disco Data Sets - Disco Constraints (5)

Disco Constraints	Data Sets: *Missy DwB*	Data Sets: *DDA-SND*
DISCO-C-VOCABULARY-01 (!)[***]		
DISCO-C-HTTP-URI-SCHEME-VIOLATION (!)[***]		

Table 25: Evaluation of Disco Data Sets - Disco Constraints (6)

4 Evaluation of Aggregated Metadata (*Data Cube*)

In this section, the quality of the metadata on aggregated data (*Data Cube*) data sets is evaluated by validating appropriate RDF constraints assigned to several RDF constraint types. First, we show the results of the evaluation of diverse data sets, then we give an overview over the evaluated data sets, and finally we provide details about the evaluation.

4.1 Evaluation Results

Table 26 shows the results of the evaluation of *Data Cube* data sets.

Evaluation Criteria	**Counts**
Validated Triples	3,775,983,610
Validated Data Sets	9,990
Constraint Violations	45,635,846
Constraint Violations ($\mathcal{SL}_0$)	0 (0%)
Constraint Violations ($\mathcal{SL}_1$)	45,520,613 (99.75%)
Constraint Violations ($\mathcal{SL}_2$)	115,233 (0.25%)
Constraint (Most Constraint Violations)	*DATA-MODEL-CONSISTENCY-05* (45,514,102)
Constraint (Most Constraint Violations ($\mathcal{SL}_0$))	-
Constraint (Most Constraint Violations ($\mathcal{SL}_1$))	*DATA-MODEL-CONSISTENCY-05* (45,514,102)
Constraint (Most Constraint Violations ($\mathcal{SL}_2$))	*MINIMUM-QUALIFIED-CARDINALITY-RESTRICTIONS-02* (1,556)
Constraint Types	20 (7|13)
Constraint Types ($\mathcal{C}_C$)	5 (25%)
Constraint Types ($\mathcal{C}_T$)	15 (75%)
Constraint Types (Most Constraints)	1. Data Model Consistency: 11 (31.4%|10|1)
	2. Existential Quantifications: 4 (11.4%|4|0)
Constraint Type (Most Constraints ($\mathcal{SL}_2$))	Data Model Consistency: 8 (22.9%|7|1)
Constraints	35 (20|15)
Constraints ($\mathcal{C}_C$)	16 (45.7%|12|4)
Constraints ($\mathcal{C}_T$)	19 (54.3%|8|11)
Constraints ($\mathcal{SL}_0$)	4 (11.4%|0|4)
Constraints ($\mathcal{SL}_1$)	3 (8.6%|3|0)
Constraints ($\mathcal{SL}_2$)	28 (80%|17|11)

Table 26: Evaluation of Data Cube Data Sets - Evaluation Results

4.2 Data Sets Overview

There are websites giving an overview over available *Data Cube* data sets[16]. Tables 27 and 29 give an overview over the evaluated *Data Cube* data sets, their abbreviations, and publicly available SPARQL endpoints. Table 28 comprehends the number of triples, data sets, and instances of multiple vocabulary-specific classes.

[16] http://270a.info/; http://datahub.io/de/dataset?tags=format-qb; http://ontologycentral.com/

Abbr.	Data Cube Data Sets
ECB	*European Central Bank*[17]
UIS	*UNESCO Institute for Statistics*[18]
IMF	*International Monetary Fund*[19]
BFS	*Bundesamt für Statistik - Swiss Federal Statistics*[20]
FAO	*Food and Agriculture Organization of the United Nations*[21]
WB	*World Bank*[22]
FRB	*Federal Reserve Board*[23]
TI	*Transparency International*[24]
OECD	*Organisation for Economic Co-operation and Development*[25]
BIS	*Bank for International Settlements*[26]
ABS	*Australian Bureau of Statistics*[27]
IEEE-VIS	*IEEE VIS Source Data*
ACORN-SAT	*Australian Climate Observations Reference Network - Surface Air Temperature Dataset*
HDP	*HealthData.gov Platform (HDP) on the Semantic Web*
Eurostat	*The Eurostat Linked Data* (SPARQL endpoint unavailable)
Asturias	*Nomenclator Asturias* (SPARQL endpoint unavailable!)
ISTAT	*ISTAT Immigration (LinkedOpenData.it)* (SPARQL endpoint unavailable)
ICANE	*Statistical Office of Cantabria (Instituto Cántabro de Estadística, ICANE)* (SPARQL endpoint unavailable)
EE-2009	*European Election Results 2009* (SPARQL endpoint unavailable)
EU-B	*Standard Eurobarometer* (SPARQL endpoint unavailable)
ECB-S	*European Central Bank Statistics (PublicData.eu)* (SPARQL endpoint unavailable)
CPV-2008	*Common Procurement Vocabulary (CPV) 2008* (SPARQL endpoint unavailable)
CPV-2003	*Common Procurement Vocabulary (CPV) 2003* (SPARQL endpoint unavailable)

Table 27: Data Cube Data Sets Abbreviations

[17] http://www.ecb.europa.eu/home/html/index.en.html
[18] http://www.uis.unesco.org/Pages/default.aspx
[19] http://www.imf.org/external/index.htm
[20] http://www.bfs.admin.ch/
[21] http://www.fao.org/home/en/
[22] http://www.worldbank.org/
[23] http://www.federalreserve.gov/
[24] http://www.transparency.org/
[25] http://www.oecd.org/
[26] http://www.bis.org/
[27] http://abs.gov.au/

	Counts				
Data Sets	**triples**	**qb:DataSet**	**qb:DataStructureDefinition**	**qb:Observation**	**qb:Slice**
ECB	468,899,474	55	46	>11,000,000	428,698
UIS	10,400,534	5	5	1,437,651	0
IMF	35,688,446	4	8	3,603,719	0
BFS	1,533,743	0	0	8	0
FAO	53,000,000	10	10	>7,100,000	0
WB	174,006,552	9,466	59	>17,000,000	0
FRB	185,266,900	49	98	>9,500,000	0
TI	52,233	6	6	3,928	0
OECD	304,995,160	136	140	>12,000,000	0
BIS	54,197,482	6	12	3,606,466	47,914
ABS	2,357,400,000	253	257	>11,000,000	0
IEEE-VIS	19,935,340	0	0	1,350	0
ACORN-SAT	98,381,319	0	4	0	0
HDP	12,226,427	0	0	0	0
Total	3,775,983,610	9,990			

Table 28: Data Cube Data Sets Overview

Data Sets	SPARQL Endpoints
ECB	http://ecb.270a.info/sparql
UIS	http://uis.270a.info/sparql
IMF	http://imf.270a.info/sparql
BFS	http://bfs.270a.info/sparql
FAO	http://fao.270a.info/sparql
WB	http://worldbank.270a.info/sparql
FRB	http://frb.270a.info/sparql
TI	http://transparency.270a.info/sparql
OECD	http://oecd.270a.info/sparql
BIS	http://bis.270a.info/sparql
ABS	http://abs.270a.info/sparql
ACORN-SAT	http://lab.environment.data.gov.au/sparql
HDP	http://healthdata.tw.rpi.edu/sparql

Table 29: Data Cube SPARQL Endpoints

4.3 Detailed Evaluation

In this sub section, we give details about the evaluation in form of diverse tables containing the number of constraint violations per evaluated data set and constraint of particular constraint types.

	Data Sets						
Data Model Consistency	*ECB*	*UIS*	*IMF*	*BFS*	*FAO*	*WB*	*FRB*
*DATA-MODEL-CONSISTENCY-01***	✓ (2)	✓	✓	✓	✓	✓	✓
*DATA-MODEL-CONSISTENCY-02****	✓ (2)	✓	✓	✓	✓	✓	✓
*DATA-MODEL-CONSISTENCY-03****	✓ (2)	✓	✓	✓	✓	✓	✓
*DATA-MODEL-CONSISTENCY-04****	✓ (6)	✓	✓	✓	✓	✓	14,372
*DATA-MODEL-CONSISTENCY-05***	1,198,352 (50)	✗	✗	✓	✗	✓	16,175,814 (42)
*DATA-MODEL-CONSISTENCY-06****	✓ (2)	✓	✓	✓	✓	✓	✓
*DATA-MODEL-CONSISTENCY-07****	✓ (9)	✓	99,091	✓	✓	✓	✓ (1)
*DATA-MODEL-CONSISTENCY-08****	✓ (2)	✓	✓	✓	✓	✓	✓
*DATA-MODEL-CONSISTENCY-09****	✓ (2)	✓	✓	✓	✓	✓	✓
*DATA-MODEL-CONSISTENCY-10**** (!)	-	-	-	-	-	-	-
*DATA-MODEL-CONSISTENCY-11***	6,511 (10)	✓	✓	✓	✓	✓	✓

Table 30: Evaluation of Data Cube Data Sets - Data Model Consistency (1)

	Data Sets						
Data Model Consistency	*TI*	*OECD*	*BIS*	*ABS*	*IEEE-VIS*	*ACORN-SAT*	*HDP*
*DATA-MODEL-CONSISTENCY-01***	✓	✓	✓	✓	✓	✓	✓
*DATA-MODEL-CONSISTENCY-02****	✓	✓	✓	✓	✓	8	✓
*DATA-MODEL-CONSISTENCY-03****	✓	✓	✓	✓	✓	✓	✓
*DATA-MODEL-CONSISTENCY-04****	✓	✓	✓	✓ (6)	✓	✓	✓
*DATA-MODEL-CONSISTENCY-05***	✓	21,142,838 (116)	✗	6,997,098 (246)	✓	✓	✓
*DATA-MODEL-CONSISTENCY-06****	✓	✓	✓	✓	✓	✓	✓
*DATA-MODEL-CONSISTENCY-07****	✓	✓	✓	✓ (8)	✓	✓	✓
*DATA-MODEL-CONSISTENCY-08****	✓	✓	✓	✓	✓	✓	✓
*DATA-MODEL-CONSISTENCY-09****	✓	✓	✓	✓	✓	✓	✓
*DATA-MODEL-CONSISTENCY-10**** (!)	-	-	-	-	-	-	-
*DATA-MODEL-CONSISTENCY-11***	✓	✓	✓	✓	✓	✓	✓

Table 31: Evaluation of Data Cube Data Sets - Data Model Consistency (2)

	Data Sets													
Existential Quantifications	*ECB*	*UIS*	*IMF*	*BFS*	*FAO*	*WB*	*FRB*	*TI*	*OECD*	*BIS*	*ABS*	*IEEE-VIS*	*ACORN-SAT*	*HDP*
*EXISTENTIAL-QUANTIFICATIONS-01****	9	✓	11	7	8	77	8	9	7	8	7	✓	✓	✓
*EXISTENTIAL-QUANTIFICATIONS-02****	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
*EXISTENTIAL-QUANTIFICATIONS-03****	✓	✓	✓	✓	✓	59	✓	6	✓	✓	✓	✓	4	✓
*EXISTENTIAL-QUANTIFICATIONS-04****	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 32: Evaluation of Data Cube Data Sets - Existential Quantifications

Cardinality Restrictions	Data Sets									
	ECB	*UIS*	*IMF*	*BFS*	*FAO*	*WB*	*FRB*	*TI*	*OECD*	*BIS*
MINIMUM-QUALIFIED-CARDINALITY-RESTRICTIONS-01 (!)[***]	-	-	-	-	-	-	-	-	-	-
MINIMUM-QUALIFIED-CARDINALITY-RESTRICTIONS-02[***]	✗	118	8	8	30	✓	30	✓	✗	12
MAXIMUM-QUALIFIED-CARDINALITY-RESTRICTIONS-01[***]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
EXACT-UNQUALIFIED-CARDINALITY-RESTRICTIONS-01[***]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
EXACT-QUALIFIED-CARDINALITY-RESTRICTIONS-02[***]	✓	✓	✓	✓	✓	1	✓	✓	✓	✓

Table 33: Evaluation of Data Cube Data Sets - Cardinality Restrictions (1)

Cardinality Restrictions	Data Sets			
	ABS	*IEEE-VIS*	*ACORN-SAT*	*HDP*
MINIMUM-QUALIFIED-CARDINALITY-RESTRICTIONS-01 (!)[***]	-	-	-	-
MINIMUM-QUALIFIED-CARDINALITY-RESTRICTIONS-02[***]	✗	1,350	✓	✓
MAXIMUM-QUALIFIED-CARDINALITY-RESTRICTIONS-01[***]	✓ (2)	✓	✓	✓
EXACT-UNQUALIFIED-CARDINALITY-RESTRICTIONS-01[***]	✓	✓	✓	✓
EXACT-QUALIFIED-CARDINALITY-RESTRICTIONS-02[***]	✓	✓	✓	✓

Table 34: Evaluation of Data Cube Data Sets - Cardinality Restrictions (2)

	Data Sets													
Structure	*ECB*	*UIS*	*IMF*	*BFS*	*FAO*	*WB*	*FRB*	*TI*	*OECD*	*BIS*	*ABS*	*IEEE-VIS*	*ACORN-SAT*	*HDP*
STRUCTURE-01***	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
STRUCTURE-02***	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 35: Evaluation of Data Cube Data Sets - Structure

	Data Sets													
Constraints	*ECB*	*UIS*	*IMF*	*BFS*	*FAO*	*WB*	*FRB*	*TI*	*OECD*	*BIS*	*ABS*	*IEEE-VIS*	*ACORN-SAT*	*HDP*
PROPERTY-DOMAIN-01 (!)***														
PROPERTY-RANGES-01 (!)***														
DISJOINT-PROPERTIES-01 (!)***														
DISJOINT-CLASSES-01 (!)***														
EQUIVALENT-PROPERTIES-01 (!)*														
UNIVERSAL-QUANTIFICATIONS-01 (!)***														
MEMBERSHIP-IN-CONTROLLED-VOCABULARIES-01 (!)***														
CONTEXT-SPECIFIC-VALID-CLASSES-01 (!)*														
CONTEXT-SPECIFIC-VALID-PROPERTIES-01 (!)*														
RECOMMENDED-PROPERTIES-01 (!)*														
VALUE-IS-VALID-FOR-DATATYPE-01 (!)***														
VOCABULARY-01 (!)***														

Table 36: Evaluation of Data Cube Data Sets - Constraints (1)

Constraints	Data Sets											
	ECB	*UIS*	*IMF*	*BFS*	*FAO*	*WB*	*FRB*	*TI*	*OECD*	*BIS*	*ABS*	*IEEE-VIS*	*ACORN-SAT*	*HDP*
*HTTP-URI-SCHEME-VIOLATION (!)***														

Table 37: Evaluation of Data Cube Data Sets - Constraints (2)

5 Evaluation of Thesauri (*SKOS*)

In this section, the quality of the metadata on thesauri (*SKOS*) is evaluated by validating appropriate RDF constraints assigned to several RDF constraint types. First, we show the results of the evaluation of diverse thesauri, then we give an overview over the evaluated thesauri, and finally we provide details about the evaluation.

5.1 Evaluation Results

Table 38 shows the results of the evaluation of *thesauri*.

Evaluation Criteria	**Counts**
Validated Triples	477,737,281
Validated Data Sets	4,178
Constraint Violations	5,540,988
Constraint Violations ($\mathcal{SL}_0$)	2,281,740 (41.2%)
Constraint Violations ($\mathcal{SL}_1$)	3,259,248 (58.8%)
Constraint Violations ($\mathcal{SL}_2$)	0 (0%)
Constraint (Most Constraint Violations)	*LANGUAGE-TAG-CARDINALITY-01* (2,508,903)
Constraint (Most Constraint Violations ($\mathcal{SL}_0$))	*LABELING-AND-DOCUMENTATION-06* (1,022,362)
Constraint (Most Constraint Violations ($\mathcal{SL}_1$))	*LANGUAGE-TAG-CARDINALITY-01* (2,508,903)
Constraint (Most Constraint Violations ($\mathcal{SL}_2$))	-
Constraint Types	14 (4|10)
Constraint Types ($\mathcal{C}_C$)	5 (35.7%)
Constraint Types ($\mathcal{C}_T$)	9 (64.3%)
Constraint Types (Most Constraints)	1. Structure: 10 (28.6%|8|2)
	2. Labeling and Documentation: 6 (17.1%|5|1)
	3. Language Tag Cardinality: 4 (11.4%|4|0)
Constraint Type (Most Constraints ($\mathcal{SL}_2$))	Structure: 1 (0|1)
Constraints	35 (17|18)
Constraints ($\mathcal{C}_C$)	21 (60%|13|8)
Constraints ($\mathcal{C}_T$)	14 (40%|4|10)
Constraints ($\mathcal{SL}_0$)	21 (60%|12|9)
Constraints ($\mathcal{SL}_1$)	5 (14.3%|5|0)
Constraints ($\mathcal{SL}_2$)	9 (25.7%|0|9)

Table 38: Evaluation of Thesauri Data Sets - Evaluation Results

5.2 Data Sets Overview

There is a website giving an overview over available *SKOS* data sets[28] and another one giving an overview over available thesauri[29]. Tables 39 and 41 give an overview over the evaluated thesauri, their abbreviations, and publicly available SPARQL endpoints. Table 40 comprehends the number of triples, data sets, and instances of multiple vocabulary-specific classes.

[28] http://datahub.io/de/dataset?tags=format-skos
[29] http://datahub.io/de/dataset?tags=thesaurus

Abbr.	Thesauri
TheSoz	*Thesaurus for the Social Sciences*[30]
STW	*Thesaurus for Economics*[31]
AGROVOC	*AGROVOC Multilingual agricultural thesaurus*[32]
UNESCO	*UNESCO Thesaurus*[33]
TGN	*The Getty Thesaurus of Geographic Names*[34]
EARTh	*Environmental Applications Reference Thesaurus*[35]
ODT	*Open Data Thesaurus*[36]
SLD	*Spanish Linguistic Datasets*[37]
SSWT	*Social Semantic Web Thesaurus*[38]
GBA-GU	*Thesaurus of the Geological Survey of Austria (GBA) - Geology Unit*[39]
GBA-GTS	*Thesaurus of the Geological Survey of Austria (GBA) - Geologic Time Scale*[40]
GBA-L	*Thesaurus of the Geological Survey of Austria (GBA) - Lithology*[41]
GBA-LU	*Thesaurus of the Geological Survey of Austria (GBA) - Lithotectonic Unit*[42]
GEMET	*GEneral Multilingual Environmental Thesaurus*[43]
EuroVoc	*EuroVoc*[44]
CECCT	*Clean Energy and Climate Change Thesaurus*[45]

Table 39: Thesauri Abbreviations

[30] http://www.ecb.europa.eu/home/html/index.en.html
[31] http://zbw.eu/stw/versions/latest/about
[32] http://202.45.139.84:10035/catalogs/fao/repositories/agrovoc
[33] http://skos.um.es/sparql/
[34] http://vocab.getty.edu/sparql
[35] http://linkeddata.ge.imati.cnr.it/resource/EARTh/
[36] http://vocabulary.semantic-web.at/PoolParty/wiki/OpenData
[37] http://linguistic.linkeddata.es
[38] http://vocabulary.semantic-web.at/PoolParty/wiki/semweb
[39] http://resource.geolba.ac.at/
[40] http://resource.geolba.ac.at/
[41] http://resource.geolba.ac.at/
[42] http://resource.geolba.ac.at/
[43] http://www.eionet.europa.eu/gemet/
[44] http://open-data.europa.eu/de/data/dataset/eurovoc
[45] http://data.reegle.info/thesaurus/guide

	Counts						
Thesauri	**triples**	**skos:ConceptScheme**	**sko:Concept**	**skos:broader**	**skos:narrower**	**skos:hasTopConcept**	**skos:inScheme**
TheSoz	439,153	1	8,426	13,705	13,706	0	48,529
STW	221,668	1	13,468	13,732	13732	7	13,180
AGROVOC	6,080,477	1	32,310	33,507	33,507	25	32,310
UNESCO	288,346	9	26,714	20,028	20,028	607	32,009
TGN	16,112,321	8	2,898,775	0	0	0	1,453,767
EARTh	9,287,364	11	295,375	288,208	93,827	479	295,376
ODT	3,290	6	108	93	93	30	0
SLD	7,629,211	0	31,195	0	0	0	0
SSWT	64,698	9	2,127	2,300	2,301	38	0
GBA-GU	25,718	3	878	1,005	1,005	14	0
GBA-GTS	7,875	3	213	208	208	5	0
GBA-L	9,317	1	249	249	249	4	0
GBA-LU	9,504	3	364	359	359	7	0
GEMET	372,889,229	3,680	414,659	62,193	21,685	30,806	409,290
EuroVoc	64,477,774	439	79,557	6,922	0	532	14,428
CECCT	191,336	3	3,419	3,761	3,762	28	0
Total	477,737,281	4,178					

Table 40: Thesauri Overview

Thesauri	SPARQL Endpoints
TheSoz	http://lod.gesis.org/thesoz/sparql
STW	http://zbw.eu/beta/sparql/stw/query
AGROVOC	http://202.45.139.84:10035/catalogs/fao/repositories/agrovoc
UNESCO	http://skos.um.es/sparql/
TGN	http://vocab.getty.edu/
EARTh	http://linkeddata.ge.imati.cnr.it:8890/sparql
ODT	http://vocabulary.semantic-web.at/PoolParty/sparql/OpenData
SLD	http://linguistic.linkeddata.es/sparql
SSWT	http://vocabulary.semantic-web.at/PoolParty/sparql/semweb
GBA-GU	http://resource.geolba.ac.at/PoolParty/sparql/GeologicUnit
GBA-GTS	http://resource.geolba.ac.at/PoolParty/sparql/GeologicTimeScale
GBA-L	http://resource.geolba.ac.at/PoolParty/sparql/lithology
GBA-LU	http://resource.geolba.ac.at/PoolParty/sparql/tectonicunit
GEMET	http://semantic.eea.europa.eu/sparql
EuroVoc	http://open-data.europa.eu/de/linked-data
CECCT	http://poolparty.reegle.info/PoolParty/sparql/glossary

Table 41: Thesauri SPARQL Endpoints

5.3 Detailed Evaluation

In this sub section, we give details about the evaluation in form of diverse tables containing the number of constraint violations per evaluated data set and constraint of particular constraint types.

	Data Sets									
Data Model Consistency	*TheSoz*	*STW*	*AGROVOC*	*TGN*	*UNESCO*	*ODT*	*SSWT*	*GBA-GU*	*GBA-GTS*	*GBA-L*	*GBA-LU*	*CECCT*
*DATA-MODEL-CONSISTENCY-01 (!)**												
*DATA-MODEL-CONSISTENCY-02 (!)**												
*DATA-MODEL-CONSISTENCY-03 (!)**												

Table 42: Thesauri Evaluation - Data Model Consistency (1)

	Data Sets
Data Model Consistency	*EARTh* *GEMET* *Euro Voc* *SLD*
*DATA-MODEL-CONSISTENCY-01 (!)**	
*DATA-MODEL-CONSISTENCY-02 (!)**	
*DATA-MODEL-CONSISTENCY-03 (!)**	

Table 43: Thesauri Evaluation - Data Model Consistency (2)

	Data Sets											
Labeling and Documentation	*TheSoz*	*STW*	*AGROVOC*	*TGN*	*UNESCO*	*ODT*	*SSWT*	*GBA-GU*	*GBA-GTS*	*GBA-L*	*GBA-LU*	*CECCT*
*LABELING-AND-DOCUMENTATION-01**	8,426	11,508	19,829	1,110	✗	36	1,475	5	2	✓	107	486
*LABELING-AND-DOCUMENTATION-02**	>1	✗	>100	287	✗	✓	✓	✓	✓	✓	✓	✓
*LABELING-AND-DOCUMENTATION-03**	✓	✓	1	14,114	✗	✓	✓	1	✓	✓	1	✓
*LABELING-AND-DOCUMENTATION-04 (!)**												
*LABELING-AND-DOCUMENTATION-05**	✓	✓	4	✓	1	2	2	1	✓	✓	✓	7
*LABELING-AND-DOCUMENTATION-06**	975,340	✓	✓	2	✓	✓	✓	✓	✓	✓	✓	✓

Table 44: Thesauri Evaluation - Labeling and Documentation (1)

	Data Sets			
Labeling and Documentation	*EARTh*	*GEMET*	*EuroVoc*	*SLD*
*LABELING-AND-DOCUMENTATION-01**	264,687	✗	54,911	31,195
*LABELING-AND-DOCUMENTATION-02**	✗	✗	✗	✓
*LABELING-AND-DOCUMENTATION-03**	2	✗	55,556	31,195
*LABELING-AND-DOCUMENTATION-04 (!)**				
*LABELING-AND-DOCUMENTATION-05**	39	✗	✗	978
*LABELING-AND-DOCUMENTATION-06**	302	46,718	✓	✓

Table 45: Thesauri Evaluation - Labeling and Documentation (2)

	Data Sets											
Structure	*TheSoz*	*STW*	*AGROVOC*	*TGN*	*UNESCO*	*ODT*	*SSWT*	*GBA-GU*	*GBA-GTS*	*GBA-L*	*GBA-LU*	*CECCT*
*STRUCTURE-01***	1	1,074	✓	✓	1	5	1	✓	✓	✓	✓	✓
*STRUCTURE-02 (!)**												
*STRUCTURE-03***	✓	✓	✓	✓	84	✓	✓	✓	✓	✓	✓	✓
*STRUCTURE-04**	2,906	8,046	726	✓	3,840	12	124	84	256	68	22	2,422
*STRUCTURE-05**	✓	✓	✓	✓	✗	90	5,150	✓	✓	✓	✓	9,864
*STRUCTURE-06**	1,457	37	✓	✓	✗	✓	4	1	1	64	✓	136
*STRUCTURE-07***	40	5,370	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓
*STRUCTURE-08 (!)****												
*STRUCTURE-09**	7,897	19,844	99	✓	552	2	16	26	✓	✓	✓	82
*STRUCTURE-10***	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 46: Thesauri Evaluation - Structure (1)

	Data Sets			
Structure	*EARTh*	*GEMET*	*Euro Voc*	*SLD*
*STRUCTURE-01***	18,240	✗	55,757	31,195
*STRUCTURE-02 (!)**				
*STRUCTURE-03***	39	4,244	✓	✓
*STRUCTURE-04**	11,286	74	✓	✓
*STRUCTURE-05**	✓	✗	✓	✓
*STRUCTURE-06**	239,346	✗	13,876	✓
*STRUCTURE-07***	110,015	✗	366,155	155,975
*STRUCTURE-08 (!)****				
*STRUCTURE-09**	107,195	32	✓	✓
*STRUCTURE-10***	27	2,122	✓	✓

Table 47: Thesauri Evaluation - Structure (2)

	Data Sets										
Language Tag Cardinality	*TheSoz*	*STW*	*AGROVOC*	*TGN UNESCO ODT*	*SSWT*	*GBA-GU*	*GBA-GTS*	*GBA-L*	*GBA-LU*	*CECCT*	
*LANGUAGE-TAG-CARDINALITY-01***	9,435	13,468	98,894	✓	541	10,147	5,117	2,061	1,742	2,272	15,550
*LANGUAGE-TAG-CARDINALITY-02**	8,222	36,936	✗	✓	265	3,627	2,212	635	631	1,253	9,607
*LANGUAGE-TAG-CARDINALITY-03**	8,222	✓	135	✓	✓	✓	✓	✓	✓	✓	✓
*LANGUAGE-TAG-CARDINALITY-04**	✓	476	✗	50	✓	✓	✓	✓	✓	✓	✓

Table 48: Thesauri Evaluation - Language Tag Cardinality (1)

Language Tag Cardinality	Data Sets: *EARTh*	*GEMET*	*EuroVoc*	*SLD*
*LANGUAGE-TAG-CARDINALITY-01***	✗	2,318,895	✗	30,781
*LANGUAGE-TAG-CARDINALITY-02**	✗	✗	✗	✗
*LANGUAGE-TAG-CARDINALITY-03**	224,206	✗	✗	31,195
*LANGUAGE-TAG-CARDINALITY-04**	✗	✗	✓	✓

Table 49: Thesauri Evaluation - Language Tag Cardinality (2)

Constraints	Data Sets: *TheSoz*	*STW*	*AGROVOC*	*TGN*	*UNESCO*	*ODT*	*SSWT*	*GBA-GU*	*GBA-GTS*	*GBA-L*	*GBA-LU*	*CECCT*
*PROPERTY-DOMAIN-01 (!)****												
*PROPERTY-RANGES-01 (!)****												
*DISJOINT-PROPERTIES-01 (!)****												
*DISJOINT-PROPERTIES-02 (!)****												
*DISJOINT-CLASSES-01 (!)****												
*EQUIVALENT-PROPERTIES-01 (!)**												
*UNIVERSAL-QUANTIFICATIONS-01 (!)****												
*CONTEXT-SPECIFIC-VALID-CLASSES-01 (!)**												
*CONTEXT-SPECIFIC-VALID-PROPERTIES-01 (!)**												
*RECOMMENDED-PROPERTIES-01 (!)**												
*VOCABULARY-01 (!)****												
*HTTP-URI-SCHEME-VIOLATION (!)****												

Table 50: Thesauri Evaluation - Constraints (1)

Constraints	Data Sets			
	EARTh	*GEMET*	*Euro Voc*	*SLD*
PROPERTY-DOMAIN-01 (!)***				
PROPERTY-RANGES-01 (!)***				
DISJOINT-PROPERTIES-01 (!)***				
DISJOINT-PROPERTIES-02 (!)***				
DISJOINT-CLASSES-01 (!)***				
EQUIVALENT-PROPERTIES-01 (!)*				
UNIVERSAL-QUANTIFICATIONS-01 (!)***				
CONTEXT-SPECIFIC-VALID-CLASSES-01 (!)*				
CONTEXT-SPECIFIC-VALID-PROPERTIES-01 (!)*				
RECOMMENDED-PROPERTIES-01 (!)*				
VOCABULARY-01 (!)***				
HTTP-URI-SCHEME-VIOLATION (!)***				

Table 51: Thesauri Evaluation - Constraints (2)

6 Evaluation of Rectangular Data (*PHDD*)

In this section, the quality of rectangular (*PHDD*) data sets is evaluated by validating appropriate RDF constraints assigned to several RDF constraint types. First, we show the results of the evaluation of diverse data sets, then we give an overview over the evaluated data sets, and finally we provide details about the evaluation.

6.1 Evaluation Results

6.2 Data Sets Overview

6.3 Detailed Evaluation

In this sub section, we give details about the evaluation in form of diverse tables containing the number of constraint violations per evaluated data set and constraint of particular constraint types.

7 Evaluation of Statistical Classifications (*XKOS*)

In this section, the quality of the metadata on statistical classifications (*XKOS*) data sets is evaluated by validating appropriate RDF constraints assigned to several RDF constraint types. First, we show the results of the evaluation of diverse data sets, then we give an overview over the evaluated data sets, and finally we provide details about the evaluation.

7.1 Evaluation Results

7.2 Data Sets Overview

Abbr.	Statistical Classifications
NAF	*Nomenclature d'activités française*[46]
PCS	*Nomenclature des Professions et Catégories Socioprofessionnelles*[47]
CJ	*Nomenclature des catégories juridiques*[48]
ISIC	
ISCO	

Table 52: Statistical Classifications Abbreviations

Nomenclature d'activités française (NAF) is the French refinement of the *NACE* classification expressed in XKOS having explanatory notes. *Nomenclature des Professions et Catégories Socioprofessionnelles (PCS)* and *Nomenclature des catégories juridiques (CJ)* are French classifications expressed in XKOS. The statistical classification *ISIC* has explanatory notes too.

7.3 Detailed Evaluation

In this sub section, we give details about the evaluation in form of diverse tables containing the number of constraint violations per evaluated data set and constraint of particular constraint types.

8 Related Work

The data most often used in research within the SBE community is *person-level data*, i.e. data collected about individuals, businesses, and households in form of responses to studies or taken from administrative registers (such as hospital records, registers of births and deaths). The range of person-level data covers

[46] http://rdf.insee.fr/codes/index.html
[47] http://rdf.insee.fr/codes/index.html
[48] http://rdf.insee.fr/codes/index.html

many different domains and is very broad - including census, education, and health data as well as all types of business, social, and labor force surveys. Increasingly, this type of research data is held within data archives or data libraries after it has been collected, so that it may be reused by future researchers. In performing their research, the detailed person-level data is aggregated into less confidential multi-dimensional tables which answer particular research questions. Portals harvest metadata (as well as publicly available data) from multiple data providers in form of RDF. To ensure high quality, the metadata must satisfy certain criteria - specified in terms of RDF constraints. After validating the metadata according to these constraints, portals offer added values to their customers, e.g. by searching over and comparing metadata of multiple providers.

By its nature, person-level data is highly confidential and access is often only permitted for qualified researchers who must apply for access. The purpose of publicly available aggregated data, on the other hand, is to get a first overview and to gain an interest in further analyses on the underlying person-level data. Researchers typically represent their results as aggregated data in form of two-dimensional tables with only a few columns (so-called *variables* such as *sex* or *age*). The *RDF Data Cube Vocabulary (QB)*[49] is a W3C recommendation for representing *data cubes*, i.e. multi-dimensional aggregate data, in RDF [4]. Aggregate data is derived from person-level data by statistics on groups or aggregates such as counts, means, and frequencies. The SDMX metadata standard – used as the basis for *QB* – and DDI have traditionally made efforts to align their content. Similarly, some of the developers of *Disco* were also involved in the development of *QB*, allowing the RDF versions of these standards to retain that alignment. While *Disco* and *QB* provide terms for the description of data sets, both on a different level of aggregation, the *Data Catalog Vocabulary (DCAT)*[50] enables the representation of these data sets inside of data collections like repositories, catalogs, or archives. The relationship between data collections and their contained data sets is useful, since such collections are a typical entry point when searching for data. Although, in most cases aggregated data is still published in form of PDFs, it is more and more common to publish aggregated data as CSV files, allowing to perform first calculations (either using all variables or only a subset). In 2014, SBE and Linked Data community members developed the *Physical Data Description (PHDD)*[51] vocabulary to represent aggregated and person-level data in a rectangular format. The data could be either represented in records with character-separated values (CSV) or in records with fixed length.

For more detailed analyses, researchers refer to person-level data from which aggregated data is derived from, as person-level data include additional variables needed for further research. One very common example for detailed analyses on person-level data is the content-driven comparison of multiple studies. Researchers get promising findings (in form of published tables with a few columns) within a metadata portal leading to subsequent research questions like 'How to

[49] http://www.w3.org/TR/vocab-data-cube/
[50] http://www.w3.org/TR/vocab-dcat/
[51] https://github.com/linked-statistics/physical-data-description

compare the unemployment rate of different countries (e.g. Germany, UK, and France) in the last 10 years grouped by age?'. The first step is to determine in which countries the unemployment rate is collected and which other variables of each country-specific study are theoretically comparable and can therefore be used to answer the underlying research question. A *study* represents the process by which a data set was generated or collected. Variables are constructed out of values (of one or multiple datatypes) and/or code lists. The variable *age*, e.g., may be represented by values of the datatype *xsd:nonNegativeInteger*, or by a code list including multiple age clusters (such as '0 to 10' and '11 to 20'). To determine if variables measuring *age* - collected within multiple studies of different countries (age_{DE}, age_{UK}) - are comparable, both content-driven and technology-driven validation is performed. An example for a content-driven validation is to investigate if variables are represented in a compatible way, i.e. are the variables' code lists theoretically comparable. Technically, it can be validated (1) if variable definitions are available, (2) if code lists are properly structured, and (3) if for each code an associated category (a human-readable label) is specified.

Data providers and harvesters do not only offer metadata but also publicly available data on different level of detail. To ensure high data quality and trust, they have to analyze and validate the data (are fundamental data fragments available?, how does valid data look like?). Provenance (where does the data come from?) is an important aspect in evaluating data quality. As data searchers know exactly which data sources they trust and which are reasonable to meet their individual use cases, RDF data validation can only be performed semi-automatically, i.e., an automatic approach serves as basis for intellectual decisions.

9 Conclusion and Future Work

We implemented a validation environment (available at http://purl.org/net/rdfval-demo) to validate RDF data according to constraints expressed my arbitrary constraint languages and to ensure correct syntax, semantics, and integrity of diverse vocabularies such as *Disco, QB, PHDD, SKOS, and XKOS*. We exhaustively evaluated the metadata quality of large real world aggregated data sets (*QB*), person-level data sets (*Disco*), thesauri (*SKOS*), statistical classifications (*XKOS*), and rectangular data sets (*PHDD*) by means of 212 $\mathcal{C}_C$ and $\mathcal{C}_T$ constraints of the majority of the constraint types. In total, we validated more than 4.2 billion triples and 15 thousand data sets.

References

1. Thomas Bosch and Kai Eckert. Towards description set profiles for rdf using sparql as intermediate language. *Proceedings of the DCMI International Conference on Dublin Core and Metadata Applications (DC 2014)*, 2014.
2. Thomas Bosch, Andreas Nolle, Erman Acar, and Kai Eckert. Rdf validation requirements - evaluation and logical underpinning. 2015.

3. Thomas Bosch, Benjamin Zapilko, Joachim Wackerow, and Kai Eckert. Rdf constraints to validate metadata on person-level, aggregated, thesauri, and statistical classifications data sets and rectangular data. 2015.
4. Richard Cyganiak, Simon Field, Arofan Gregory, Wolfgang Halb, and Jeni Tennison. Semantic statistics: Bringing together sdmx and scovo. In Christian Bizer, Tom Heath, Tim Berners-Lee, and Michael Hausenblas, editors, *Proceedings of the WWW 2010 Workshop on Linked Data on the Web*, volume 628 of *CEUR Workshop Proceedings*, 2010.