An Evaluation of RDF Constraint Types to Validate Metadata on Person-Level and Aggregated Data

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. ...

Keywords: RDF Validation, RDF Constraints, DDI-RDF Discovery Vocabulary, Disco, RDF Data Cube Vocabulary, Linked Data, Semantic Web

1 Introduction

2 Evaluation

2.1 Legend

Symbol	Description
X	**Validation Successful**
$> X$	**Poor Performance/Scaling**
✗	**Very Poor Performance/Scaling**
†	**SPARQL Endpoint Error**

Table 1: Legend

Validation Successful: 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 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 Disco

– DwB Discovery Portal

- Metadata records for each study are available on the study landing pages.
- http://dwb-dev.nsd.uib.no/portal/#/study/http://portal.dwbproject.org/resource/Study/de.gesis/zacat/ZA5597/en
- Search interface: http://dwb-dev.nsd.uib.no/portal/#/search?q=smoking&p=1&s=score&d=desc&f=text

Abbr.	**Disco Data Sets**
Missy	Microdata Information System[3]
DwB	DwB Discovery Portal[4]
DDA-SND	DDA and SND DDI-RDF[5]

Table 2: Disco Data Sets Abbreviations

[3] http://www.gesis.org/missy/eu/missy-home
[4] http://dwb-dev.nsd.uib.no/portal
[5] http://ddi-rdf.borsna.se/

	Counts									
Data Sets	**triples**	**disco:StudyGroup**	**disco:Study**	**disco:LogicalDataSet**	**disco:Universe**	**disco:Variable**	**disco:Question**	**disco:SummaryStatistics**	**disco:CategoryStatistics**	**skos:Concept**
Missy										
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

Table 3: Disco Data Sets Overview

Data Sets	**SPARQL Endpoint**
Missy	XXXXX
DwB	http://dwb-dev.nsd.uib.no/sparql
DDA-SND	http://ddi-rdf.borsna.se/endpoint/

Table 4: Disco SPARQL Endpoints

4 Data Cube

– overview over Data Cube data sets[6]

– http://ontologycentral.com/

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

Abbr.	Data Cube Data Sets
ECB	European Central Bank[7]
UIS	UNESCO Institute for Statistics[8]
IMF	International Monetary Fund[9]
BFS	Bundesamt für Statistik - Swiss Federal Statistics[10]
FAO	Food and Agriculture Organization of the United Nations[11]
WB	World Bank[12]
FRB	Federal Reserve Board[13]
TI	Transparency International[14]
OECD	Organisation for Economic Co-operation and Development[15]
BIS	Bank for International Settlements[16]
ABS	Australian Bureau of Statistics[17]
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
Asturias	Nomenclator Asturias
ISTAT	ISTAT Immigration (LinkedOpenData.it)
ICANE	Statistical Office of Cantabria (Instituto Cántabro de Estadística, ICANE)
EE-2009	European Election Results 2009
EU-B	Standard Eurobarometer
ECB-S	European Central Bank Statistics (PublicData.eu)
CPV-2008	Common Procurement Vocabulary (CPV) 2008
CPV-2003	Common Procurement Vocabulary (CPV) 2003
GHI-2011	Global Hunger Index (GHI) 2011

Table 5: Data Cube Data Sets Abbreviations

[7] http://www.ecb.europa.eu/home/html/index.en.html
[8] http://www.uis.unesco.org/Pages/default.aspx
[9] http://www.imf.org/external/index.htm
[10] http://www.bfs.admin.ch/
[11] http://www.fao.org/home/en/
[12] http://www.worldbank.org/
[13] http://www.federalreserve.gov/
[14] http://www.transparency.org/
[15] http://www.oecd.org/
[16] http://www.bis.org/
[17] 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	61,406,503				
HDP	1,934,046,908				
Eurostat	8,000,000,000				
Asturias	4,508,050				
ISTAT	3,024,396				
ICANE	52,412				
EE-2009	3,165				
EU-B	1,193,494				
ECB-S	10,000,000				
CPV-2008	803,311				
CPV-2003	546,135				
GHI-2011	19,369 (dump)				

Table 6: 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

Table 7: Data Cube Data Sets SPARQL Endpoints

	Data Sets											
Data Model Consistency	ECB	UIS	IMF	BFS	FAO	WB	FRB	TI	OECD	BIS	ABS	IEEE-VIS
DATA-MODEL-CONSISTENCY-01	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
DATA-MODEL-CONSISTENCY-02	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
DATA-MODEL-CONSISTENCY-03	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
DATA-MODEL-CONSISTENCY-04	✗	✓	✓	✓				✓				
DATA-MODEL-CONSISTENCY-05												✓
DATA-MODEL-CONSISTENCY-06												
DATA-MODEL-CONSISTENCY-07												
DATA-MODEL-CONSISTENCY-08												
DATA-MODEL-CONSISTENCY-09												
DATA-MODEL-CONSISTENCY-10												
DATA-MODEL-CONSISTENCY-11												

Table 8: Evaluation of Data Cube Data Sets

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

Table 9: Evaluation of Data Cube Data Sets

	Data Sets											
Cardinality Restrictions	ECB	UIS	IMF	BFS	FAO	WB	FRB	TI	OECD	BIS	ABS	IEEE-VIS
MINIMUM-QUALIFIED-CARDINALITY-RESTRICTIONS-01												
EXACT-UNQUALIFIED-CARDINALITY-RESTRICTIONS-01	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
EXACT-QUALIFIED-CARDINALITY-RESTRICTIONS-01	✗	118	8	8	✗	✗	✗	✓	✗	12	✗	1,350
EXACT-QUALIFIED-CARDINALITY-RESTRICTIONS-02	✓	✓	✓	✓	✓	1	✓	✓	✓	✓	✓	✓

Table 10: Evaluation of Data Cube Data Sets

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

Table 11: Evaluation of Data Cube Data Sets

Constraints	Data Sets									
	ECB	UIS	IMF	BFS	FAO	WB	FRB	TI	OECD	BIS	ABS	IEEE-VIS
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 12: Evaluation of Data Cube Data Sets

5 SKOS

- overview over SKOS data sets[18]
- overview over thesauri[19] (table 14)

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

Abbr.	Thesauri
TheSoz	Thesaurus for the Social Sciences[20]
STW	Thesaurus for Economics[21]
AGROVOC	AGROVOC Multilingual agricultural thesaurus[22]
UNESCO	UNESCO Thesaurus[23]
TGN	The Getty Thesaurus of Geographic Names[24]
EARTh	Environmental Applications Reference Thesaurus[25]
ODT	Open Data Thesaurus[26]
SLD	Spanish Linguistic Datasets[27]
SSWT	Social Semantic Web Thesaurus[28]
GBA-GU	Thesaurus of the Geological Survey of Austria (GBA) - Geology Unit[29]
GBA-GTS	Thesaurus of the Geological Survey of Austria (GBA) - Geologic Time Scale[30]
GBA-L	Thesaurus of the Geological Survey of Austria (GBA) - Lithology[31]
GBA-LU	Thesaurus of the Geological Survey of Austria (GBA) - Lithotectonic Unit[32]
GEMET	GEneral Multilingual Environmental Thesaurus[33]
EuroVoc	EuroVoc[34]
CECCT	Clean Energy and Climate Change Thesaurus[35]

Table 13: Thesauri Abbreviations

[20] http://www.ecb.europa.eu/home/html/index.en.html
[21] http://zbw.eu/stw/versions/latest/about
[22] http://202.45.139.84:10035/catalogs/fao/repositories/agrovoc
[23] http://skos.um.es/sparql/
[24] http://vocab.getty.edu/sparql
[25] http://linkeddata.ge.imati.cnr.it/resource/EARTh/
[26] http://vocabulary.semantic-web.at/PoolParty/wiki/OpenData
[27] http://linguistic.linkeddata.es
[28] http://vocabulary.semantic-web.at/PoolParty/wiki/semweb
[29] http://resource.geolba.ac.at/
[30] http://resource.geolba.ac.at/
[31] http://resource.geolba.ac.at/
[32] http://resource.geolba.ac.at/
[33] http://www.eionet.europa.eu/gemet/
[34] http://open-data.europa.eu/de/data/dataset/eurovoc
[35] 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						
ODT	3,290						
SLD	7,629,211						
SSWT	64,698						
GBA-GU	25,718						
GBA-GTS	7,875						
GBA-L	9,317						
GBA-LU	9,504						
GEMET	372,889,229						
EuroVoc	64,477,774						
CECCT	191,336						

Table 14: 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 15: Thesauri SPARQL Endpoints

	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 16: Thesauri Evaluation

	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	✓	36	1,475	6	2	✓	107	486
LABELING-AND-DOCUMENTATION-02	>1	✗	>100	✓	✓	✓	✓	✓	✓	✓	✓
LABELING-AND-DOCUMENTATION-03	✓	✓	1	✓	✓	✓	1	✓	✓	1	✓
LABELING-AND-DOCUMENTATION-04											
LABELING-AND-DOCUMENTATION-05	✓	✓	✗	✓	✓	✓	1	✓	✓	✓	3
LABELING-AND-DOCUMENTATION-06	>1	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓

Table 17: Thesauri Evaluation

	Data Sets										
Structure	TheSoz	STW	AGROVOC	TGN UNESCO ODT	SSWT	GBA-GU	GBA-GTS	GBA-L	GBA-LU	CECCT
STRUCTURE-01	1	1,074	✓	✓	5	1	✓	✓	✓	✓	✓
STRUCTURE-02											
STRUCTURE-03	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
STRUCTURE-04	2,906	>1,000	726	✓	12	124	84	256	>54	22	2,422
STRUCTURE-05	✓	✓	✓	✓	90	5,150	✓	✓	✓	✓	9,864
STRUCTURE-06	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	113
STRUCTURE-07	8	1,074	✓	✓	✓	✓	✓	✓	✓	✓	✓
STRUCTURE-08											
STRUCTURE-09	6,741	>3,297	135	✓	2	16	25	✓	✓	✓	81
STRUCTURE-10	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 18: Thesauri Evaluation

	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 19: Thesauri Evaluation

	Data Sets										
Constraints	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												

Table 20: Thesauri Evaluation

6 XKOS

- Nomenclature d'activités française (NAF): French classification expressed in XKOS. the French refinement of the NACE, because it has explanatory notes.
- Nomenclature des Professions et Catégories Socioprofessionnelles (PCS): French classification expressed in XKOS.
- Nomenclature des catégories juridiques (CJ): French classification expressed in XKOS.
- ISIC: has explanatory notes too.
- ISCO

	Counts
Data Sets	**triples**
Nomenclature d'activités française (NAF)[36]	
Nomenclature des Professions et Catégories Socioprofessionnelles (PCS)[37]	
Nomenclature des catégories juridiques (CJ)[38]	
ISIC	
ISCO	

Table 21: Statistical Classifications Overview

References

[36] http://rdf.insee.fr/codes/index.html
[37] http://rdf.insee.fr/codes/index.html
[38] http://rdf.insee.fr/codes/index.html