Rensselaer
why not change the world?®


Tetherless World Constellation

SciKG Part 3: Semantic Data Dictionaries (SDD)

Henrique Santos, Paulo Pinheiro, Jamie P. McCusker, Sabbir M. Rashid, Deborah L. McGuinness
May 28th 2023

The 20th Extended Semantic Web Conference (ESWC-23)

The 20th Extended Semantic Web Conference (ESWC-23)
May 28th, 2023

SciKG Part 3: Semantic Data Dictionaries (SDDs)

Henrique Santos Paulo Pinheiro Jamie P. McCusker
Sabbir M. Rashid Deborah L. McGuinness

Columbia University
MAILMAN SCHOOL
OF PUBLIC HEALTH

TABLE OF CONTENTS

01

INTRODUCTION

We introduce this work, motivate its importance, and present our claims

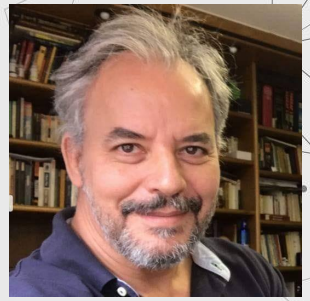
Data Int. journal paper authors - https://bit.ly/3kG6iDi

SABBIR RASHID

JAMIE MCCUSKER

PAULO PINHEIRO

MARCELLO BAX

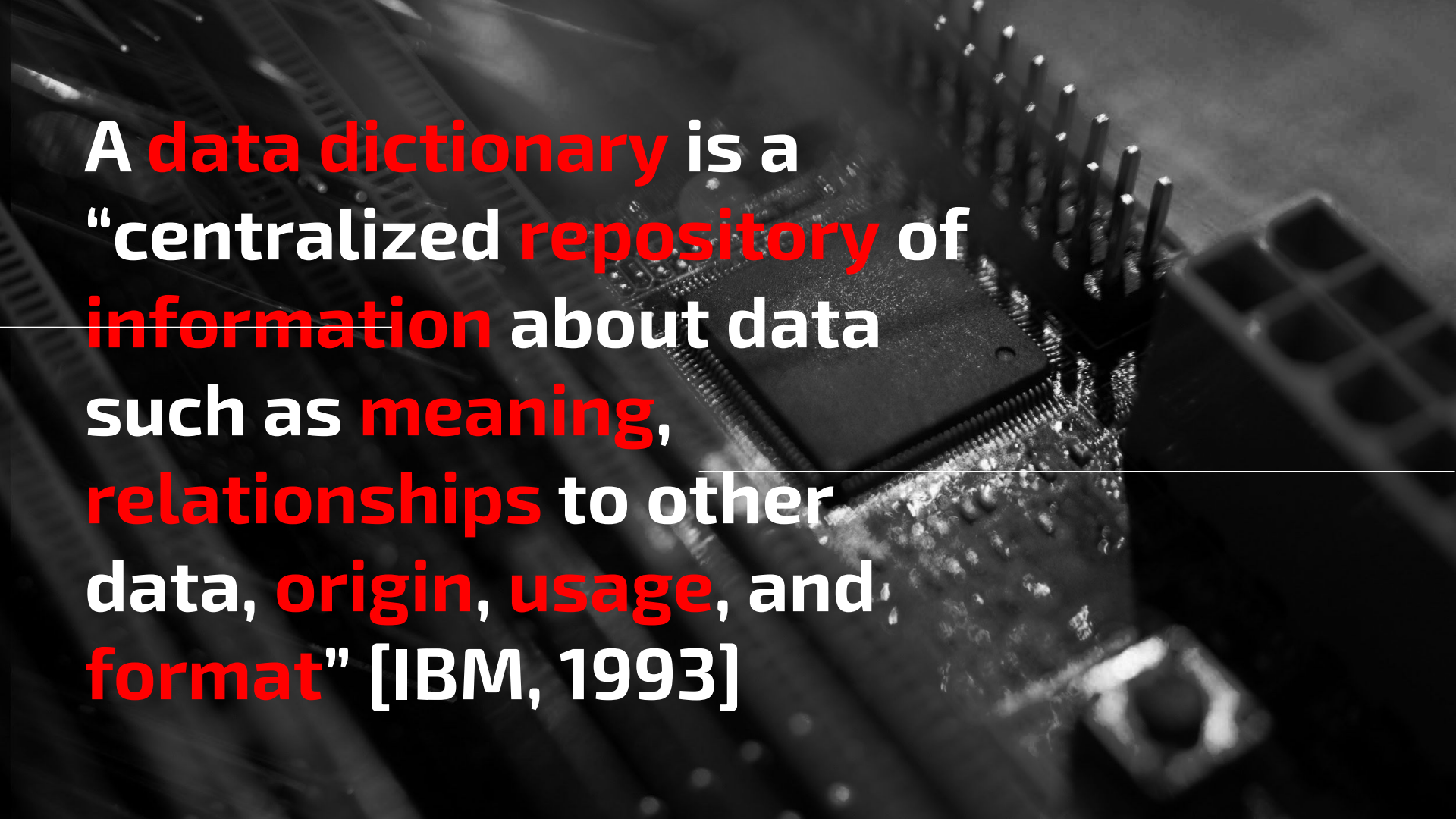


HENRIQUE SANTOS

JEANETTE STINGONE

AMAR DAS

DEBORAH MCGUINNESS


A data dictionary is a
“centralized repository of
information about data
such as meaning,
relationships to other
data, origin, usage, and
format” [IBM, 1993]

DATA DICTIONARIES

- Ambiguity
- Standard adherence
- Human consumption

LIMITATIONS

- Semantic technology usage
- Implicit concept annotation
- Provenance incorporation

SEMANTIC IMPROVEMENTS

MOTIVATION

- Annotate data from various domains
- Harmonize data from multiple sources
- Understand the data

CLAIMS

ADDRESSES LIMITATIONS

Addresses limitations of traditional data dictionaries

Semantic Data Dictionary

An approach for annotating and transforming data

ABSTRACTION

Presents a level of abstraction over mapping language-based approaches

F.A.I.R.

Resulting model is Findable, Accessible, Interoperable, and Reusable

02

RELATED WORK

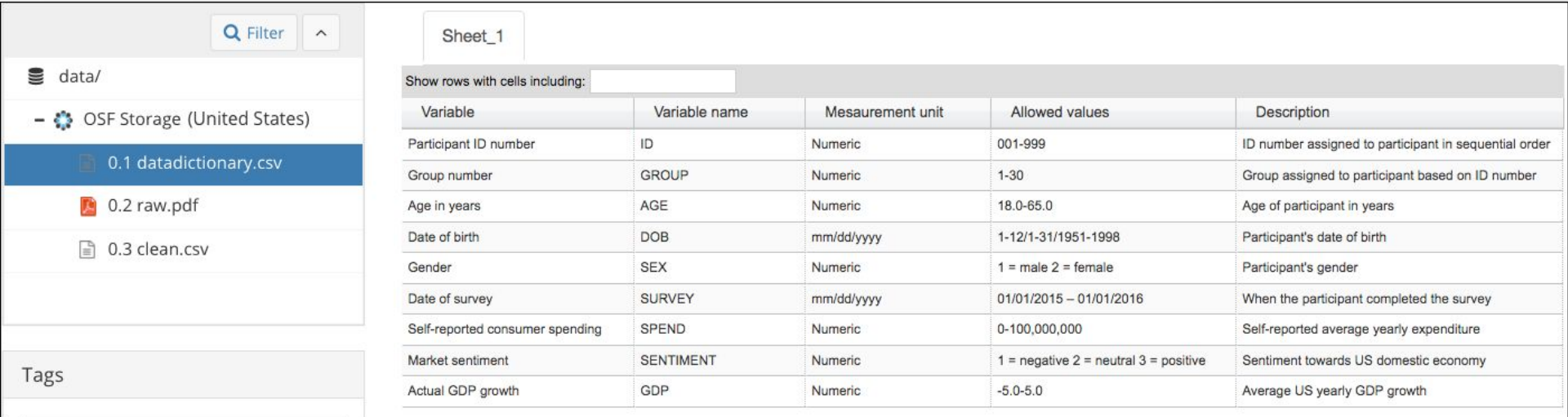
We review literature related to traditional data dictionaries, data integration, mapping languages, and semantic ETL

TRADITIONAL DATA DICTIONARIES

- Data dictionaries mentioned in patents [Haskell et al., 2009, Lau et al., 2002, Apacible et al., 2013]
- Stony Brook Data Governance Council (https://bit.ly/3oD4q90)
- The Open Science Framework (https://bit.ly/35EPupT)
- Biosystematic Database of World Diptera [Thompson, 1999]
- Project Open Data Metadata Schema (https://bit.ly/3oCYTqr)

LIMATIONS

- Minimal incorporation of semantics
- Object and relation elicitation not permitted
- Domain-specific
- Not machine-readable
- Lack of a formal creation standard



Variable	Variable name	Mesaurement unit	Allowed values	Description
Participant ID number	ID	Numeric	001-999	ID number assigned to participant in sequential order
Group number	GROUP	Numeric	1-30	Group assigned to participant based on ID number
Age in years	AGE	Numeric	18.0-65.0	Age of participant in years
Date of birth	DOB	mm/dd/yyyy	1-12/1-31/1951-1998	Participant's date of birth
Gender	SEX	Numeric	1 = male 2 = female	Participant's gender
Date of survey	SURVEY	mm/dd/yyyy	01/01/2015 – 01/01/2016	When the participant completed the survey
Self-reported consumer spending	SPEND	Numeric	0-100,000,000	Self-reported average yearly expenditure
Market sentiment	SENTIMENT	Numeric	1 = negative 2 = neutral 3 = positive	Sentiment towards US domestic economy
Actual GDP growth	GDP	Numeric	-5.0-5.0	Average US yearly GDP growth

DATA INTEGRATION

- Techniques that utilize data from multiple sources to construct a unified view of the combined data [Lenzerini, 2002]

- The Semantic Web Integration Tool (SWIT) [del Carmen Legaz-García et al., 2016]
- RDF-Gen [Santipantakis et al., 2018]
- DataOps [Pinkel et al., 2015]
- OpenRefine [Ham, 2013]

LIMATIONS

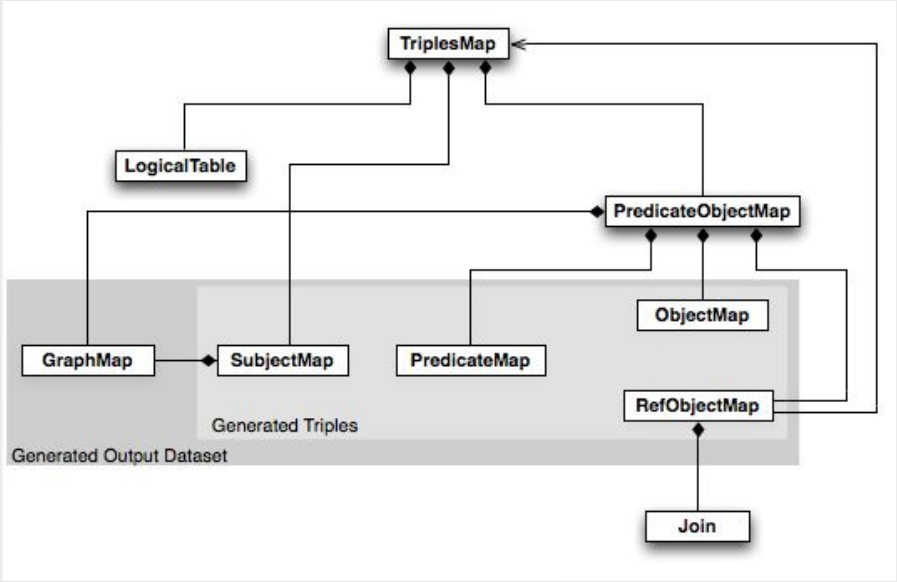
- Not all tools are open source
- Some require knowledge of mapping languages
- Difficulties with subset selection, cell-based operations, dataset merging
- Not all tools allow object elicitation
- Some difficulties associated with adoption

MAPPING LANGUAGES

- Typically used to convert a relational database (RDB), tabular file, or hierarchical structure to an RDF format

---

- RDB to RDF Mapping Language (R2RML) [Arenas et al., 2012]
- RDF Mapping Language (RML) [Dimou et al., 2014]
- xR2RML [Michel et al., 2015]
- KR2RML [Slepicka et al., 2015]
- Karma [Knoblock and Szekely, 2015]
- Sparqlification Mapping Language (SML) [Stadler et al., 2015]
- RDB2OWL [Cerans and Bumans, 2011]



https://www.w3.org/TR/r2rml/

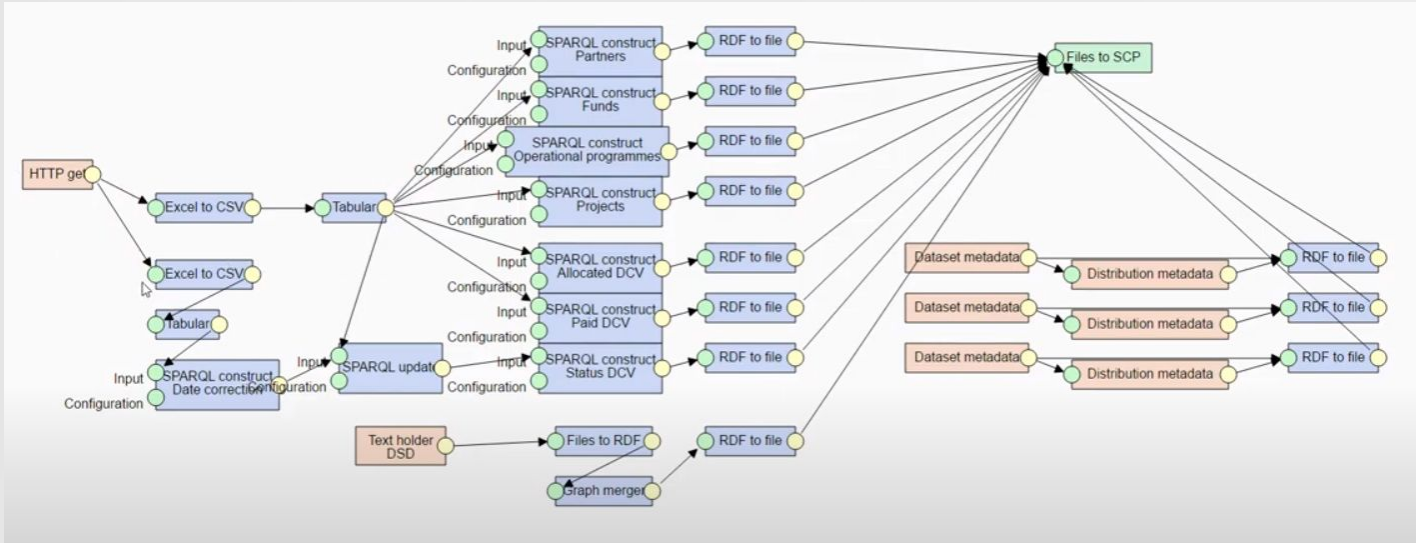
SEMANTIC EXTRACT, TRANSFORM, AND LOAD (ETL)

- (ETL) operations refer to processes that read data from a source database, convert the data into another format, and write the data into a target database

---

- LinkedPipes ETL (LP-ETL) [Klımek et al., 2016]
- Semantic Extract, Transform, and Load-er (SETLr) [McCusker et al., 2018]
- Eureka! Clinical Analytics [Post et al., 2013]
- Linked Data Integration Framework (LDIF) [Schultz et al., 2011]

https://etl.linkedpipes.com/



03

SEMANTIC DATA DICTIONARY

We present the various components included in the Semantic Data Dictionary specification

SEMANTIC DATA DICTIONARY SPECIFICATION

INFOSHEET

Contains links to the other specifications

DICTIONARY MAPPING

Used to annotate the columns of a dataset

CODEBOOK

Used to annotate coded values

METADATA SUPPLEMENT

Includes metadata about the Semantic Data Dictionary or the associated dataset

CODE MAPPING

Used to encode shortcut notations (See https://bit.ly/2HLydmK)

TIMELINE

Used for complex temporal mappings

PROPERTIES

Used to customize the properties used during the mapping process

INFOSHEET

SPECIFICATION

Infosheet Row	Description
Code Mapping	Reference to Code Mapping table location
Codebook	Reference to Codebook table location
Dictionary Mapping	Reference to Dictionary Mapping table location
Properties	Reference to Properties table location
Timeline	Reference to Timeline table location

METADATA SUPPLEMENT

Infosheet Row	Related Property	Description
Contributors	*dct:contributor*	Contributors to the SDD
Creators	*dct:creator*	Creators of the SDD
Date Created	*dct:created*	Date the SDD was created
Description	*dct:description*	Description of the KG fragment
Imports	*owl:imports*	Ontologies that the SDD references
Keywords	*schema:keywords*	Keywords to be associated with the KG fragment
License	*dct:license*	License URL
Previous Version	*pav:previousVersion*	Previous version URL
Publisher	*dct:publisher*	Publisher of the SDD
Title	*dct:title*	Title of KG fragment
Version	*owl:versionInfo*	Current version URL
Version Of	*dct:isVersionOf*	Resource URL for primary version

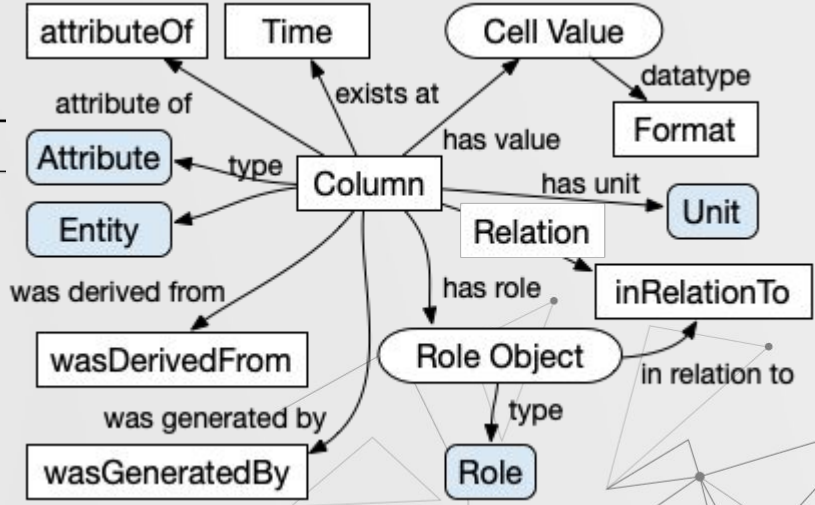
- Data on the Web Best Practices
  - https://www.w3.org/TR/dwbp/
- Semantic Web Health Care and Life Sciences
  - https://www.w3.org/TR/hcls-dataset/

DICTIONARY MAPPING

SPECIFICATION

DM Column	Related Property	Description
Attribute	*rdf:type*	Class of attribute entry
attributeOf	*sio:isAttributeOf*	Entity having the attribute
Column		Entry column header in dataset
Entity	*rdf:type*	Class of entity entry
Format		Specifies the structure of the cell value
inRelationTo	*sio:inRelationTo*	Entity that the role is linked to
Label	*rdfs:label*	Label for the entry
Relation		Custom property used in inRelationTo
Role	*sio:hasRole*	Type of the role of the entry
Time	*sio:existsAt*	Time point of measurement
Unit	*sio:hasUnit*	Unit of measure for entry
wasDerivedFrom	*prov:wasDerivedFrom*	Entity from which the entry was derived
wasGeneratedBy	*prov:wasGeneratedBy*	Activity from which the entry was produced

DIAGRAM



DICTIONARY MAPPING FORMALISM

$$\exists \text{Column} \wedge \exists \text{Attribute} \implies \text{Attribute}(\text{Column})$$

$$\exists \text{Column} \wedge \exists \text{Entity} \implies \text{Entity}(\text{Column})$$

$$\exists \text{Column} \wedge \exists \text{Label} \implies \textit{rdfs:label}(\text{Column}, \text{Label})$$

$$\exists \text{Column} \wedge \exists \text{Comment} \implies \textit{rdfs:comment}(\text{Column}, \text{Comment})$$

$$\exists \text{Column} \wedge \exists \text{Definition} \implies \textit{skos:definition}(\text{Column}, \text{Definition})$$

$$\exists \text{Column} \wedge \exists \text{attributeOf} \implies \textit{sio:attributeOf}(\text{Column}, \text{attributeOf})$$

$$\exists \text{Column} \wedge \exists \text{Unit} \implies \exists \text{U} \wedge \text{Unit}(\text{U}) \wedge \textit{sio:hasUnit}(\text{Column}, \text{U})$$

$$\exists \text{Column} \wedge \exists \text{Format} \wedge \exists \text{Value} \implies \textit{sio:hasValue}(\text{Column}, \text{Value}\hat{}\hat{}\text{Format})$$

$$\exists \text{Column} \wedge \exists \text{Time} \implies \textit{sio:existsAt}(\text{Column}, \text{Time})$$

$$\exists \text{Column} \wedge \exists \text{Role} \implies \exists \text{R} \wedge \textit{sio:hasRole}(\text{Column}, \text{R}) \wedge \text{Role}(\text{R})$$

$$\exists \text{Column} \wedge \exists \text{Role} \wedge \exists \text{inRelationTo} \implies \exists \text{R} \wedge \textit{sio:hasRole}(\text{Column}, \text{R}) \wedge \text{Role}(\text{R}) \wedge \textit{sio:inRelationTo}(\text{R}, \text{inRelationTo})$$

$$\exists \text{Column} \wedge \exists \text{inRelationTo} \implies \textit{sio:inRelationTo}(\text{Column}, \text{inRelationTo})$$

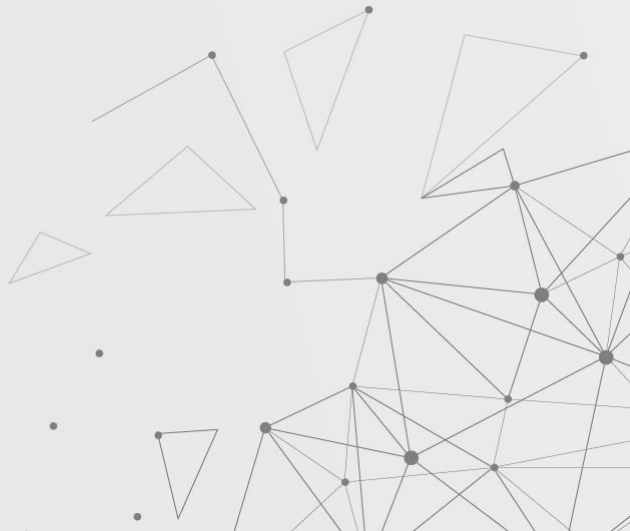
$$\exists \text{Column} \wedge \exists \text{Relation} \wedge \exists \text{inRelationTo} \implies \text{Relation}(\text{Column}, \text{inRelationTo})$$

$$\exists \text{Column} \wedge \exists \text{Role} \wedge \exists \text{Relation} \wedge \exists \text{inRelationTo} \implies \exists \text{R} \wedge \textit{sio:hasRole}(\text{Column}, \textit{R}) \wedge \text{Role}(\text{R}) \wedge \text{Relation}(\text{R}, \text{inRelationTo})$$

$$\exists \text{Column} \wedge \exists \text{wasDerivedFrom} \implies \textit{prov:wasDerivedFrom}(\text{Column}, \text{wasDerivedFrom})$$

$$\exists \text{Column} \wedge \exists \text{wasGeneratedBy} \implies \textit{prov:wasGeneratedBy}(\text{Column}, \text{wasGeneratedBy})$$

$$\exists \text{Column} \wedge \exists \text{Value} \implies \textit{sio:hasValue}(\text{Column}, \text{Value})$$

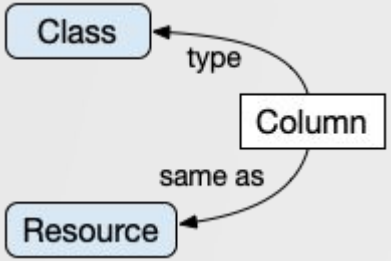


CODEBOOK

SPECIFICATION

Codebook Column	Related Property	Description
CLASS	*rdf:type*	Class the Code refers to
CODE	*sio:hasValue*	Value of the dataset entry
COLUMN		Entry column header in dataset
LABEL	*rdfs:label*	Label for the codebook entry
RESOURCE	*rdf:type*	Web Resource URI the Code refers to

DIAGRAM



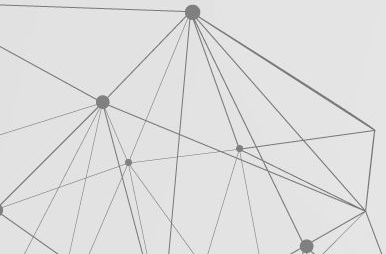
FORMALISM

$$\exists \text{COLUMN} \wedge \exists \text{CLASS} \implies \text{CLASS}(\text{COLUMN})$$

$$\exists \text{COLUMN} \wedge \exists \text{LABEL} \implies \textit{rdfs:label}(\text{COLUMN}, \text{LABEL})$$

$$\exists \text{COLUMN} \wedge \exists \text{RESOURCE} \implies \textit{owl:sameAs}(\text{COLUMN}, \text{RESOURCE})$$

$$\exists \text{COLUMN} \wedge \exists \text{CODE} \implies \textit{sio:hasValue}(\text{COLUMN}, \text{CODE})$$



TIMELINE

SPECIFICATION

Timeline Column	Related Property	Description
END	*sio:hasEndTime*	The starting time point associated with the Timeline entry
INRELATIONTO	*sio:inRelationTo*	Entity that the Timeline entry is associated with
NAME		Implicit entry reference for the Timeline entry
START	*sio:hasStartTime*	The starting time point associated with the Timeline entry
TYPE	*rdf:type*	Class the Timeline entry refers to
UNIT	*sio:hasUnit*	Unit of measure for Timeline entry

FORMALISM

$$\exists \text{NAME} \wedge \exists \text{TYPE} \implies \text{TYPE}(\text{NAME})$$

$$\exists \text{NAME} \wedge \exists \text{LABEL} \implies \textit{rdfs:label}(\text{NAME})$$

$$\exists \text{NAME} \wedge \exists \text{START} \implies \exists \text{S} \wedge \textit{sio:hasStartTime}(\text{NAME}, \text{S}) \wedge \textit{sio:hasValue}(\text{S}, \text{START})$$

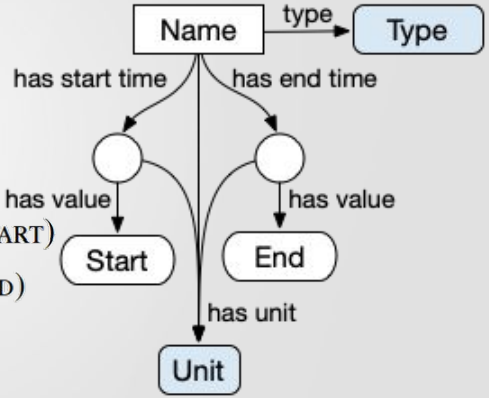
$$\exists \text{NAME} \wedge \exists \text{END} \implies \exists \text{E} \wedge \textit{sio:hasEndTime}(\text{NAME}, \text{E}) \wedge \textit{sio:hasValue}(\text{E}, \text{END})$$

$$\exists \text{NAME} \wedge \exists \text{START} \wedge \exists \text{END} \wedge \text{START} \equiv \text{END} \implies \exists \text{T} \wedge \textit{sio:existsAt}(\text{NAME}, \text{T}) \wedge \textit{sio:hasValue}(\text{T}, \text{START})$$

$$\exists \text{NAME} \wedge \exists \text{UNIT} \implies \exists \text{U} \wedge \text{UNIT}(\text{U}) \wedge \textit{sio:hasUnit}(\text{NAME}, \text{U})$$

$$\exists \text{NAME} \wedge \exists \text{INRELATIONTO} \implies \textit{sio:inRelationTo}(\text{NAME}, \text{INRELATIONTO})$$

DIAGRAM



PROPERTIES SPECIFICATION

Row	Property
Attribute	*rdf:type*
attributeOf	*sio:isAttributeOf*
Comment	*rdfs:comment*
Definition	*skos:definition*
End	*sio:hasEndTime*
Entity	*rdf:type*
inRelationTo	*sio:inRelationTo*
Label	*rdfs:label*
Role	*sio:hasRole*
Start	*sio:hasStartTime*
Time	*sio:existsAt*
Type	*rdf:type*
Unit	*sio:hasUnit*
Value	*sio:hasValue*
wasDerivedFrom	*prov:wasDerivedFrom*
wasGeneratedBy	*prov:wasGeneratedBy*

04

MODELLING APPROACHES

We discuss some modeling strategies and provide some examples to help illustrate this work

SEMANTIC SCIENTIFIC WORKFLOW

ONTOLOGY ENGINEERING

CLASS SELECTION

- Collect relevant ontologies
  - http://www.ontobee.org/
  - https://bioportal.bioontology.org/

SUPPORTING ONTOLOGY

- What if concepts used to annotate dataset does not exist in an ontology?
  - Create concept map
  - Engineer a supporting ontology
    - Protege
      - https://protege.stanford.edu/
  - Manual vs. automated approaches

KNOWLEDGE GRAPH GENERATION

- https://github.com/tetherless-world/SemanticDataDictionary
  - sdd2rdf
  - sdd2setl
- https://github.com/tetherless-world/whyis

- Will cover this in the tutorial!

- Loading of KG into a triplestore
- Querying the resulting Graph

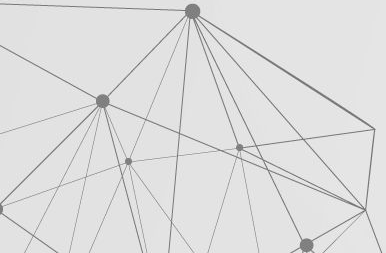
INFOSHEET EXAMPLE

Attribute	Value
Creators	Sabbir M. Rashid
Code Mapping	NHANES/config/code_mappings.csv
Codebook	NHANES/input/CB/DEMO_H_Doc-CB.csv
Contributors	"James P. McCusker, Paulo Pinheiro, Marcello P. Bax, Henrique O. Santos, Alexander New, Shruthi Chari, Mathew Johnson,John S. Erickson, Kristin P. Bennett, Jeanette A. Stingone, Deborah L. McGuinness"
Date Created	2018-10-14
Description	KG fragment from manually annotated NHANES Demographics SDD.
Dictionary Mapping	NHANES/input/DM/DEMO_H_Doc-DM.csv
Imports	"http://semanticscience.org/ontology/sio-subset-labels.owl , http://hadatac.org/ont/chear/ , http://purl.obolibrary.org/obo/ncit.owl"
Keywords	"demographics, gender, age, race, citizenship, marital status, household"
License	https://opensource.org/licenses/MIT
Previous Version	http://tw.rpi.edu/heals/kb/nhanes/1.1
Properties	NHANES/config/Properties.csv
Publisher	Tetherless World Constellation
Timeline	NHANES/input/TL/DEMO_H_Doc-TL.csv
Title	The National Health and Nutrition Examination (NHANES) SDD KG
Version	http://tw.rpi.edu/heals/kb/nhanes/1.2
Version Of	http://tw.rpi.edu/heals/kb/nhanes/

DICTIONARY MAPPING EXAMPLE

Column	Label	Attribute	attributeOf	Unit	Time	Entity	Relation	inRelationTo
SEQN	Respondent sequence number	sio:Identifier	??participant					
RIAGENDR	Gender	sio:BiologicalSex	??participant					
RIDAGEYR	Age in years at screening	sio:Age	??participant	yr	??screening			
RIDAGEMN	Age in months at screening	sio:Age	??participant	mth	??screening			
RIDRETH1	Race/Hispanic origin	sio:Race	??participant					
RIDEXAGM	Age in months at exam	sio:Age	??participant	mth	??exam			
DMDBORN4	Country of birth				??birth	sio:Country	*sio:isLocationOf*	??participant
DMDCITZN	Citizenship status	sio:StatusDescriptor	??participant					
DMDYRSUS	Length of time in US	sio:TimeInterval	??participant					
DMDEDUC3	Education level - Children/Youth	chear:EducationLevel	??participant					
DMDEDUC2	Education level - Adults 20+	chear:EducationLevel	??participant					
DMDMARTL	Marital status	chear:MaritalStatus	??participant					
RIDEXPRG	Pregnancy status at exam	sio:StatusDescriptor	??pregnancy		??exam			??participant
SIALANG	Language of SP Interview	chear:Language	??instrument		??interview			??participant
DMDHRGND	HH ref person's gender	sio:BiologicalSex	??HHRef					
DMDHRAGE	HH ref person's age in years	sio:Age	??HHRef	yr				
DMDHRBR4	HH ref person's country of birth				??birth	sio:Country	*sio:isLocationOf*	??HHRef
DMDHREDU	HH ref person's education level	chear:EducationLevel	??HHRef					
DMDHRMAR	HH ref person's marital status	chear:MaritalStatus	??HHRef					
WTINT2YR	Full sample 2 year interview wt	chear:Weight	??participant		??interview			
WTMEC2YR	Full sample 2 year MEC exam wt	chear:Weight	??participant		??exam			
INDHHIN2	Annual household income	chear:Income	??household					

EXPLICIT ENTRIES



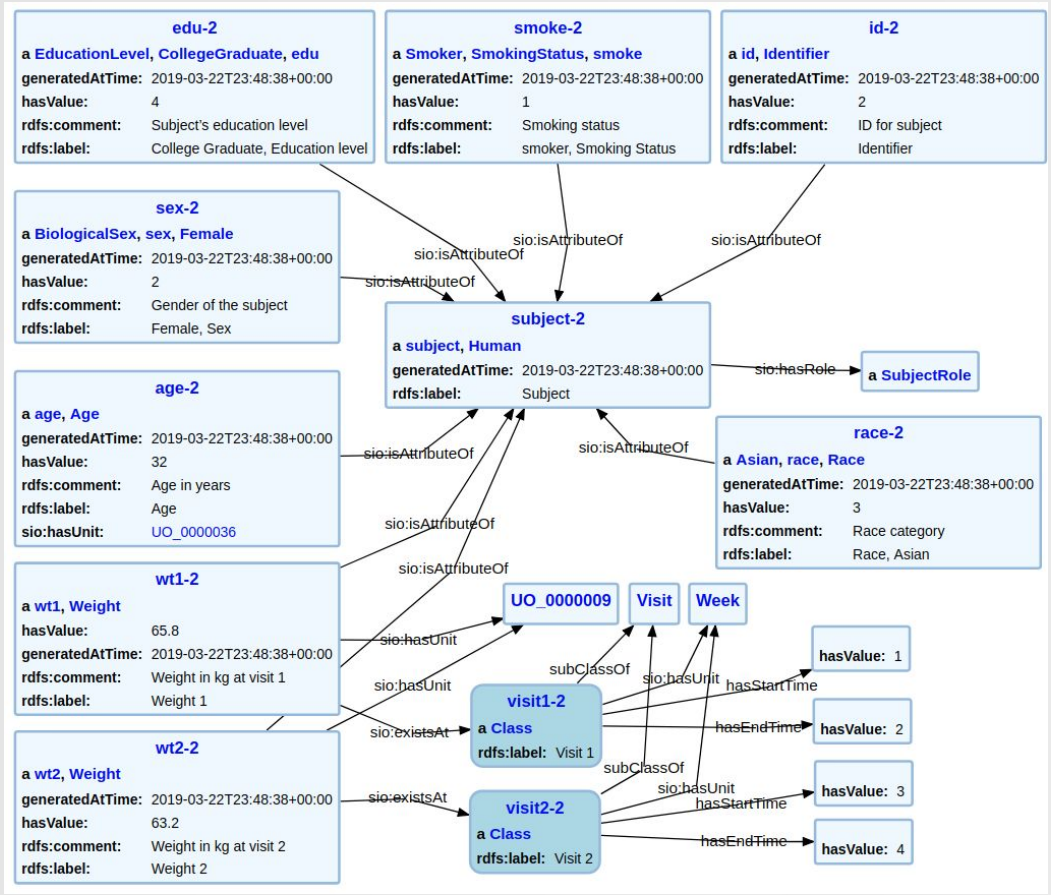
IMPLICIT ENTRIES

Column	Label	Entity	Role	inRelationTo
??participant	Participant	ncit:C29867, sio:Human	sio:SubjectRole	
??screening	Screening	chear:Screening		
??exam	Examination	ncit:C131902		
??birth	Birth	sio:Birthing		
??pregnancy	Pregnancy	chear:Pregnancy		
??interview	Interview	ncit:C16751		
??instrument	Instrumentation	ncit:C16742		
??household	Household	chear:Household		??participant
??HHRef	Household reference	sio:Human	chear:HeadOfHousehold	??household

CODEBOOK EXAMPLE

Column	Code	Label	Class
RIAGENDR	1	Male	sio:Male
RIAGENDR	2	Female	sio:Female
RIAGENDR	.	Missing	ncit:C142610
RIDRETH1	1	Mexican American	exo:0000151
RIDRETH1	2	Other Hispanic	exo:0000145
RIDRETH1	3	Non-Hispanic White	exo:0000158
RIDRETH1	4	Non-Hispanic Black	exo:0000132
RIDRETH1	5	Other Race - Including Multi-Racial	exo:0000153
RIDRETH1	.	Missing	ncit:C142610
DMDEDUC3	0	Never attended / kindergarten only	chear:NoFormalEducation
DMDEDUC3	1	1st grade	chear:EducationGrade
DMDEDUC3	2	2nd grade	chear:EducationGrade
DMDEDUC3	3	3rd grade	chear:EducationGrade
DMDEDUC3	4	4th grade	chear:EducationGrade
DMDEDUC3	5	5th grade	chear:EducationGrade
DMDEDUC3	6	6th grade	chear:EducationGrade
DMDEDUC3	7	7th grade	chear:EducationGrade
DMDEDUC3	8	8th grade	chear:EducationGrade

REPRESENTATION

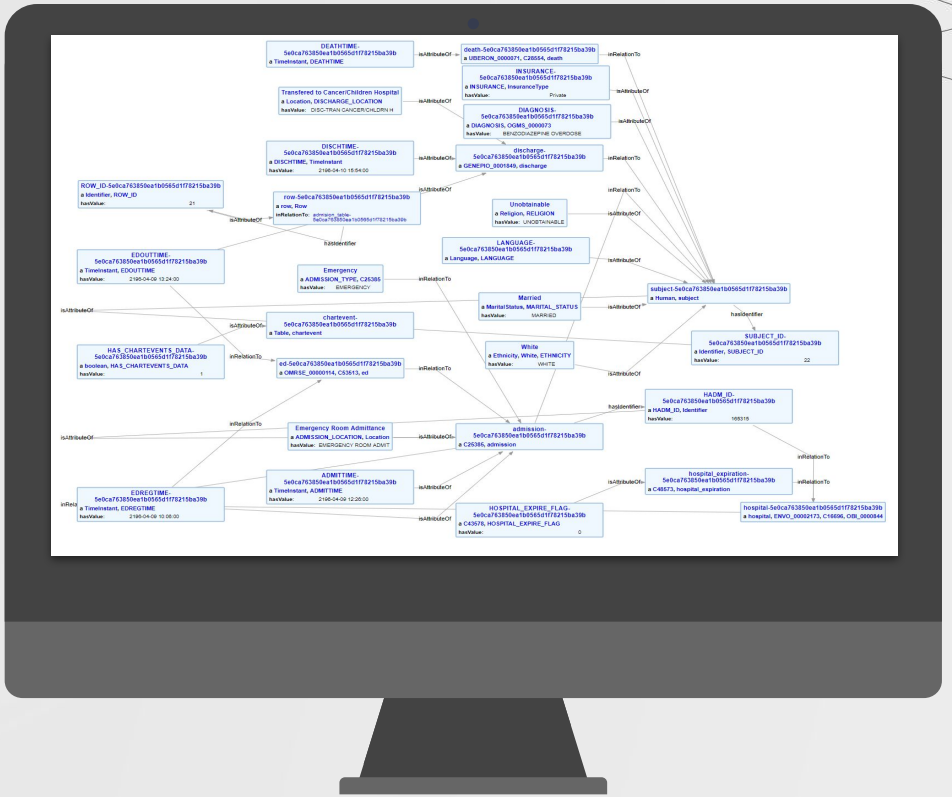


REPRESENTATION

Column	Attribute	attributeOf	Entity	Role	inRelationTo
SUBJECT_ID	sio:Identifier	??subject			
ADMITTIME	sio:TimeInstant	??admission			
DISCHTIME	sio:TimeInstant	??discharge			
DEATHTIME	sio:TimeInstant	??death			
INSURANCE	chear:InsuranceType	??subject			
RELIGION	chear:Religion	??subject			
MARITAL_STATUS	chear:MaritalStatus	??subject			
ETHNICITY	sio:Ethnicity	??subject			
DIAGNOSIS	ogms:0000073	??subject			
??subject			sio:Human	sio:SubjectRole	
??admission			ncit:C25385		??subject
??discharge			genepio:0001849		??subject
??death			ncit:C28554		??subject

```
mimic-kb:ROW_ID-5e0ca763850ea1b0565d1f78215ba39b        rdf:type        mimic-kb:ROW_ID ;
        rdf:type        sio:Identifier ;
        sio:isAttributeOf mimic-kb:row-5e0ca763850ea1b0565d1f78215ba39b ;
        sio:hasValue    "21"^^xsd:integer .

mimic-kb:SUBJECT_ID-5e0ca763850ea1b0565d1f78215ba39b    rdf:type        mimic-kb:SUBJECT_ID ;
        rdf:type        sio:Identifier ;
        sio:isAttributeOf mimic-kb:subject-5e0ca763850ea1b0565d1f78215ba39b ;
        sio:hasValue    "22"^^xsd:integer .

mimic-kb:HADM_ID-5e0ca763850ea1b0565d1f78215ba39b       rdf:type        mimic-kb:HADM_ID ;
        rdf:type        sio:Identifier ;
        sio:isAttributeOf mimic-kb:admission-5e0ca763850ea1b0565d1f78215ba39b ;
        sio:inRelationTo        mimic-kb:hospital-5e0ca763850ea1b0565d1f78215ba39b ;
        sio:hasValue    "165315"^^xsd:integer .

mimic-kb:ADMITTIME-5e0ca763850ea1b0565d1f78215ba39b     rdf:type        mimic-kb:ADMITTIME ;
        rdf:type        sio:TimeInstant ;
        sio:isAttributeOf mimic-kb:admission-5e0ca763850ea1b0565d1f78215ba39b ;
        sio:hasValue    "2196-04-09 12:26:00"^^xsd:string .

mimic-kb:DISCHTIME-5e0ca763850ea1b0565d1f78215ba39b     rdf:type        mimic-kb:DISCHTIME ;
        rdf:type        sio:TimeInstant ;
        sio:isAttributeOf mimic-kb:discharge-5e0ca763850ea1b0565d1f78215ba39b ;
        sio:hasValue    "2196-04-10 15:54:00"^^xsd:string .

mimic-kb:DEATHTIME-5e0ca763850ea1b0565d1f78215ba39b     rdf:type        mimic-kb:DEATHTIME ;
        rdf:type        sio:TimeInstant ;
        sio:isAttributeOf mimic-kb:death-5e0ca763850ea1b0565d1f78215ba39b .

mimic-kb:ADMISSION_TYPE-5e0ca763850ea1b0565d1f78215ba39b        rdf:type        mimic-kb:ADMI
        rdf:type        ncit:C25385 ;
        sio:inRelationTo        mimic-kb:admission-5e0ca763850ea1b0565d1f78215ba39b ;
        rdfs:label      "Emergency"^^xsd:string ;
        sio:hasValue    "EMERGENCY"^^xsd:string .

mimic-kb:ADMISSION_LOCATION-5e0ca763850ea1b0565d1f78215ba39b    rdf:type        mimic-kb:ADMI
        rdf:type        sio:Location ;
        sio:isAttributeOf mimic-kb:admission-5e0ca763850ea1b0565d1f78215ba39b ;
        rdfs:label      "Emergency Room Admittance "^^xsd:string ;
        sio:hasValue    "EMERGENCY ROOM ADMIT"^^xsd:string .
```

http://cs.rpi.edu/~rashis2/ont/mimic/mimic-adm.ttl



https://bit.ly/2HJ3iHD

05

CHALLENGES

We discuss some challenges faced by domain scientists when creating their of Semantic Data Dictionaries

EXPERIMENTAL SETUP

- Domain scientists were presented with initial training
  - Epidemiologists and biostatisticians
- Supporting materials were developed in collaboration with a domain expert
  - Were made available to provide guidance and examples
- A template for completing the Semantic Data Dictionary was provided
  - Included pre-populated fields for common demographic concepts
    - Such as age, race, and gender
- A help document was created that included instructions and representations of more complex concepts
  - Measurements of environmental samples
  - Measurements of biological samples
  - Measurements taken at specific time-points
- A practical workshop was held
  - A semantic scientist provided training in semantic representation to the domain scientists
- Domain scientists completed at least one Semantic Data Dictionary for an epidemiologic study

CHALLENGES

- Domain scientists had representation difficulties
  - Complex ideas (e.g. fasting blood glucose levels)
  - Implicit concepts
    - Uncommon representation in the public health domain
    - Not necessarily intuitive
  - Time associations
- Determining best ontology term for annotation was not always clear
  - What if a term was not found in a supporting ontology?
    - Best way to represent concept in a semantically appropriate way
    - What other ontologies should be used?
- Requirement for user to have some domain & ontology knowledge
- Currently only supports annotation of tabular data
- Annotation process is mostly manual
- Documentation and tutorials can be improved

06

CONCLUSIONS

Thanks for listening!

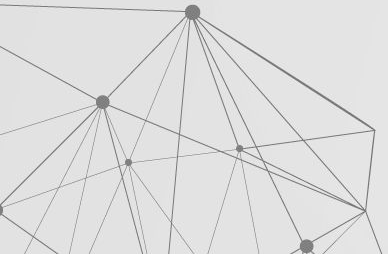
CLOSING COMMENTS

Semantic Data Dictionaries address many of the limitations of the prior work -- thus, this work helps advance the state-of-the-art

The SDD approach follows Semantic Web standards and results in artifacts that are findable, accessible, interoperable, and reusable

RESOURCES

- whyis - https://github.com/tetherless-world/whyis
- HADatAc - https://github.com/paulopinheiro1234/hadatac
- sdd2rdf - https://github.com/tetherless-world/SemanticDataDictionary
- Documentation - https://tetherless-world.github.io/sdd/
- Annotated resource examples - https://github.com/tetherless-world/sdd/tree/master/sdd_resources
- Journal Paper - https://bit.ly/3kG6iDi



REFERENCES

- Apacible, J. T., Nolan, S. P., Kalmady, G. D., and Varadan, V. (2013). Extensible and localizable health-related dictionary. US Patent 8,417,537.
- Arenas, M., Bertails, A., Prudhommeaux, E., and Sequeda, J. (2012). A direct mapping of relational data to rdf. W3C recommendation, 27:1–11.
- Cerans, K. and Bumans, G. (2011). Rdb2owl: a rdb-to-rdf/owl mapping specification language.Information Systems, pages 139–152.
- del Carmen Legaz-Garc´ıa, M., Minarro-Gimenez, J. A., Menarguez-Tortosa, M., and Fern´andez-Breis, J. T. (2016). Generation of open biomedical datasets through ontology-driven transformation and integration processes. Journal of biomedical semantics, 7(1):32.
- Dimou, A., Vander Sande, M., Colpaert, P., De Vocht, L., Verborgh, R., Mannens, E., and Van de Walle, R. (2014). Extraction and semantic annotation of workshop proceedings in html using rml. In Semantic Web Evaluation Challenge, pages 114–119. Springer.
- Ham, K. (2013). Openrefine (version 2.5). http://openrefine.org. free, open-source tool for cleaning and transforming data. Journal of the Medical Library Association: JMLA, 101(3): 233.
- Haskell, R. E., Heil, J. A., and Cassidy, J. (2009). Dynamic dictionary and term repository system.US Patent 7,580,831.
- IBM (1993). IBM Dictionary of Computing.McGraw-Hill, Inc., New York, NY, USA, 10th edition.
- Klımek, J.,Skoda, P., and Necasky, M. (2016). Linkedpipes etl: Evolved linked data preparation.In European Semantic Web Conference, pages 95–100. Springer.
- Knoblock, C. A. and Szekely, P. (2015). Exploiting semantics for big data integration.AI Magazine, 36(1).
- Kuchinke, W., Wiegelmann, S., Verplancke, P., and Ohmann, C. (2006). Extended cooperation in clinical studies through exchange of cdisc metadata between different study software solutions. Methods of information in medicine, 45(04):441–446.

REFERENCES

- Lau, L., Endo, J., Karren, S., Willis, M., Harada, S., Beeney, S., Larsen, B., Cassin, E., and Gerard, M. (2002). Mapping clinical data with a health data dictionary.US Patent App. 09/755,966.
- Lenzerini, M. (2002). Data integration: A theoretical perspective. In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pages 233–246. ACM.
- McCusker, J. P., Chastain, K., Rashid, S., Norris, S., and McGuinness, D. L. (2018). Setlr: the semantic extract, transform, and load-r.PeerJ Preprints, 6:e26476v1.
- Michel, F., Djimenou, L., Zucker, C. F., and Montagnat, J. (2015).Translation of relational and non-relational databases into rdf with xr2rml.In11th International Conference on Web Information Systems and Technologies (WEBIST'15), pages 443–454.
- Pinkel, C., Schwarte, A., Trame, J., Nikolov, A., Bastinos, A. S., and Zeuch, T. (2015). Dataops: seamless end-to-end anything-to-rdf data integration.In European Semantic Web Conference, pages 123–127. Springer.
- Post, A. R., Krc, T., Rathod, H., Agravat, S., Mansour, M., Torian, W., and Saltz, J. H. (2013). Semantic etl into i2b2 with eureka!AMIA Summits on Translational Science Proceedings, 2013:203.
- Santipantakis, G. M., Kotis, K. I., Vouros, G. A., and Doulkeridis, C. (2018). Rdf-gen: Generating rdf from streaming and archival data.In Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, page 28. ACM.
- Schultz, A., Matteini, A., Isele, R., Bizer, C., and Becker, C. (2011). Ldif-linked data integration framework.InProceedings of the Second International Conference on Consuming Linked Data-Volume 782, pages 125–130.CEUR-WS. Org.
- Slepicka, J., Yin, C., Szekely, P. A., and Knoblock, C. A. (2015). Kr2rml: An alternative interpretation of r2rml for heterogenous sources. In Cold.

REFERENCES

- Stadler, C., Unbehauen, J., Westphal, P., Sherif, M. A., and Lehmann, J. (2015). Simplified rdb2rdf mapping.In LDOW@ WWW, volume 1409.
- Thompson, F. C. (1999). Data dictionary and standards for fruit fly information database.
- Myia.Warzel, D. B., Andonyadis, C., McCurry, B., Chilukuri, R., Ishmukhamedov, S., and Covitz, P. (2003). Common data element (cde) management and deployment in clinical trials.InAMIA annual symposium proceedings, volume 2003, page 1048. American Medical Informatics Association.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W.,da Silva Santos, L. B., Bourne, P. E., et al. (2016).The fair guiding principles for scientific data management and stewardship.Scientific data, 3.
- Zozus, M. N., Bonner, J., and Rock, L. (2017).Towards data value-level metadata for clinical studies. In ITCH, pages 418–423.