

The background of the slide is a light gray gradient with several realistic water droplets of various sizes scattered across it. The droplets have highlights and shadows, giving them a three-dimensional appearance.

ShEx in Reference Quality and Subsetting

SEYED AMIR HOSSEINI BEGHAEIRAVERI

Entity Schemas In The Wikimedia Ecosystem Tutorial

SWAT4HCLS 2022

DATA QUALITY

Quality is a measure of "fitness for use"

Data quality is a **MULTI-DIMENSIONAL** concept

Dimensions of the data quality:

- Availability
- Believability
- Completeness
- Relevancy
- Free-of-Error
- ...

Most of dimensions are **SUBJECTIVE**

DATA QUALITY IN LINKED DATA [1]

Category	Dimensions
Accessibility	Availability, Licensing, Interlinking, Security, Performance
Intrinsic	Accuracy, Consistency, Conciseness
Trust	Reputation, Believability, Verifiability, Objectivity
Dynamicity	Currency, Volatility, Timeliness
Contextual	Completeness, Amount-of-data, Relevancy
Representational	Representational-conciseness, Representational-consistency, Understandability, Interpretability, Versatility

QUALITY OF REFERENCING

6

CATEGORIES

21

DIMENSIONS

40

METRICS

4

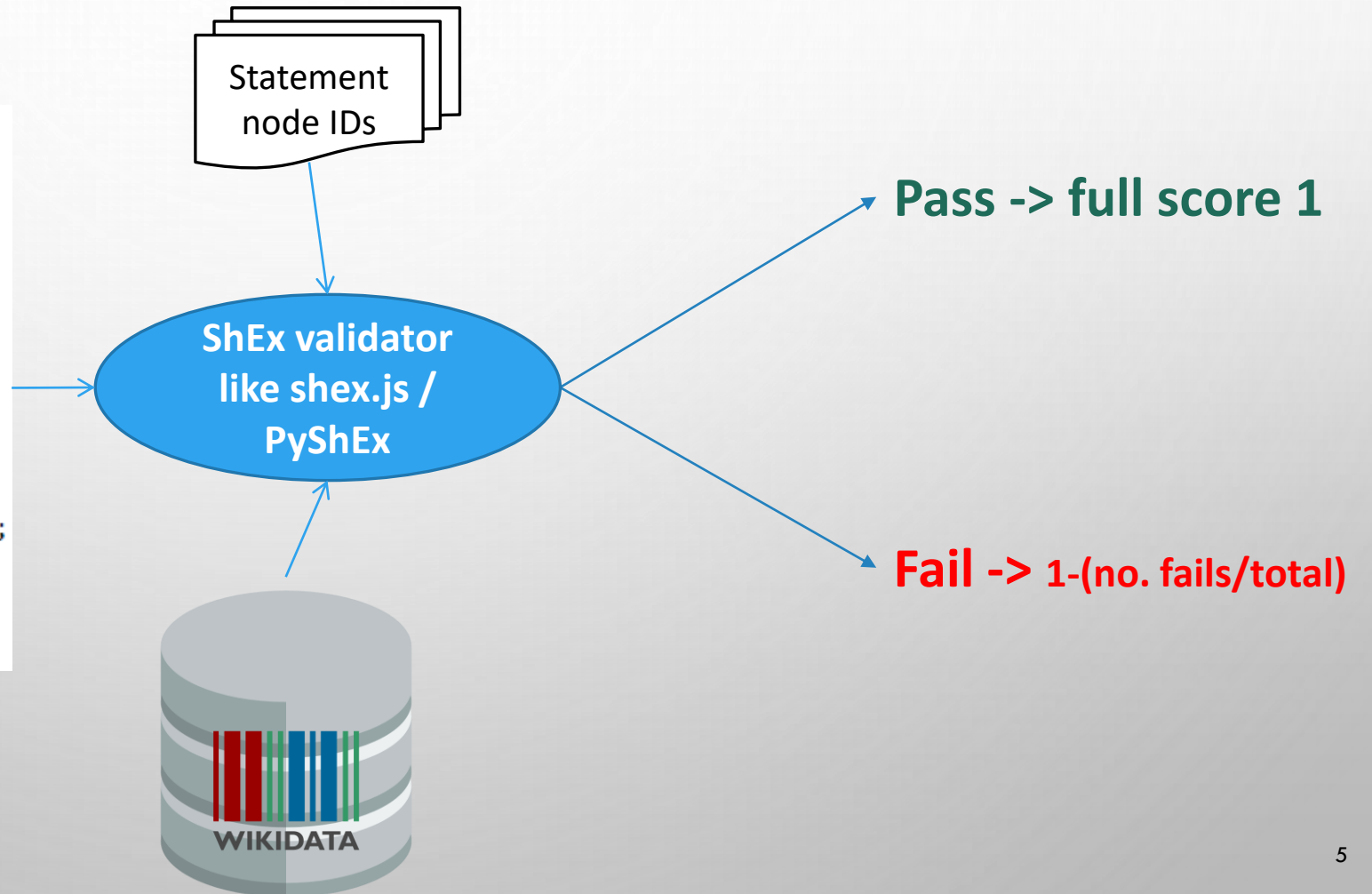
SHEX-RELATED
METRICS

DIMENSION: ACCURACY

Syntactic validity of reference triples

```
START=@<statement_node>
<statement_node> {
  prov:wasDerivedFrom @<reference_node>* ;
}
<reference_node> {
  <ref-property> xsd:decimal OR xsd:integer OR
  xsd:dateTime OR xsd:string OR IRI* ;
  <ref-property> @<value>* ;
}
<value> {
  wikibase:quantityAmount xsd:decimal OR xsd:integer?;
  wikibase:timeValue xsd:dateTime?;
  #...
}
```

Wikidata RDF model
of references Shape



DIMENSION: COMPLETENESS

Schema completeness of references

How many classes and properties have a defined Entity Schemas (Eids) for references?

```
<#reference> { } # Any reference will suffice.

<#disease-ontology-reference> { # reference to a term from the disease ontology term
  pr:P248 [ wd:Q5282129 ] ; # stated in [P248] Mondo disease ontology [Q27468140]
  pr:P699 xsd:string ; # Disease Ontology ID
  pr:P813 xsd:dateTime ; # Date of retrieval
}

<#mondo-disease-reference> { # reference to a term from the MonDo ontology
  pr:P248 [ wd:Q27468140 ] ; # stated in [P248] Mondo disease ontology [Q27468140]
  pr:P5270 xsd:string ; # Mondo ID
  pr:P813 xsd:dateTime ; # Date of retrieval
}

<#symptom-ontology-reference> { # reference to a term from the Symptom Ontology
  pr:P248 [ wd:Q5282129 ] ; # stated in [P248] Symptom ontology [Q27468140]
  pr:P8656 xsd:string ; # Symptom Ontology ID
  pr:P813 xsd:dateTime ; # Date of retrieval
}
```

DIMENSION: COMPLETENESS

Schema completeness of references

Gene Wiki disease terms (E113)	disease (E69)	virus strain (E170)	virus gene (E165)	protein (E167)	virus protein (E169)	virus gene (E165) Circular reference
				virus protein (E169)	virus gene (E165) Circular reference	
		virus protein (E169)	virus gene (E165)	protein (E167)	virus protein (E169) Circular reference	
				virus protein (E169) Circular reference		
Gene Wiki primary sources (E273)	Gene Wiki disease terms (E113)	virus strain (E170)	virus gene (E165)	protein (E167)	virus protein (E169)	virus gene (E165) Circular reference
				virus protein (E169)	virus gene (E165) Circular reference	
		virus protein (E169)	virus gene (E165)	protein (E167)	virus protein (E169) Circular reference	
				virus protein (E169) Circular reference		
Gene Wiki external identifiers (E274)	Gene Wiki disease terms (E113)	virus strain (E170)	virus gene (E165)	protein (E167)	virus protein (E169)	virus gene (E165) Circular reference
				virus protein (E169)	virus gene (E165) Circular reference	
		virus protein (E169)	virus gene (E165)	protein (E167)	virus protein (E169) Circular reference	
				virus protein (E169) Circular reference		
Gene Wiki symptom terms (E275)	Gene Wiki symptom terms (E275)	virus taxon (E192)	virus strain (E170)	virus gene (E165)	protein (E167)	virus protein (E169) Circular reference
				virus protein (E169)	virus gene (E165) Circular reference	
		virus strain (E170)	virus gene (E165)	protein (E167)	virus protein (E169) Circular reference	
				virus protein (E169) Circular reference		
https://www.wikidata.org/wiki/Wikidata:Database_reports/EntitySchema_directory						

DIMENSION: COMPLETENESS

Schema-based property completeness of references

If a reference schema is defined for class/fact of type X, how many instances of X have got a reference with the property mentioned in the schema?

```
<#reference> { } # Any reference will suffice.

<#disease-ontology-reference> { # reference to a term from the disease ontology term
  pr:P248 [ wd:Q5282129 ] ; # stated in [P248] Mondo disease ontology [Q27468140]
  pr:P699 xsd:string ; # Disease Ontology ID
  pr:P813 xsd:dateTime ; # Date of retrieval
}

<#mondo-disease-reference> { # reference to a term from the Mondo ontology
  pr:P248 [ wd:Q27468140 ] ; # stated in [P248] Mondo disease ontology [Q27468140]
  pr:P5270 xsd:string ; # Mondo ID
  pr:P813 xsd:dateTime ; # Date of retrieval
}

<#symptom-ontology-reference> { # reference to a term from the Symptom Ontology
  pr:P248 [ wd:Q5282129 ] ; # stated in [P248] Symptom ontology [Q27468140]
  pr:P8656 xsd:string ; # Symptom Ontology ID
  pr:P813 xsd:dateTime ; # Date of retrieval
}
```

How many disease has got reference using P248/P699/P813

=>

How many symptoms has got reference using P248/P8656/P813

...

DIMENSION: COMPLETENESS

Property completeness of references

If a fact of type X has a reference using the ref-property Y, how many other type X facts have a reference using property Y?

Albert Einstein (Q937)
German-born theoretical physicist; developer

<u>relative</u> X	<ul style="list-style-type: none">Lina Einstein kinship to subject ▼ 1 reference <u>reference URL</u> YElsa Einstein kinship to subject ▼ 1 reference
languages spoken, written or signed	<ul style="list-style-type: none">German ▼ 0 references no Y!

Ernest Rutherford (Q9123)
New Zealand-born British chemist and physicist (1871-1937)

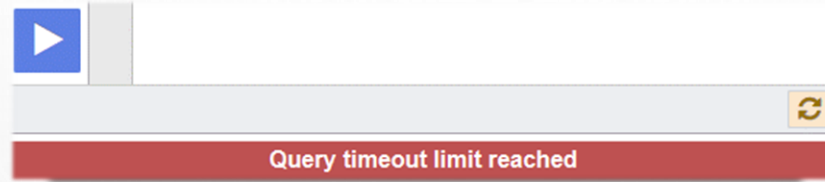
<u>relative</u> X	<ul style="list-style-type: none">Ralph H. Fowler kinship to subject ▼ 0 references
languages spoken, written or signed	<ul style="list-style-type: none">English no Y! ▼ 2 references stated in Bibliothèque nationale de France ID reference URL retrieved reference URL Y

SUBSETTING KGS



Huge Size of KGs

Wikidata 2021
100 GB



Timed-Out Queries

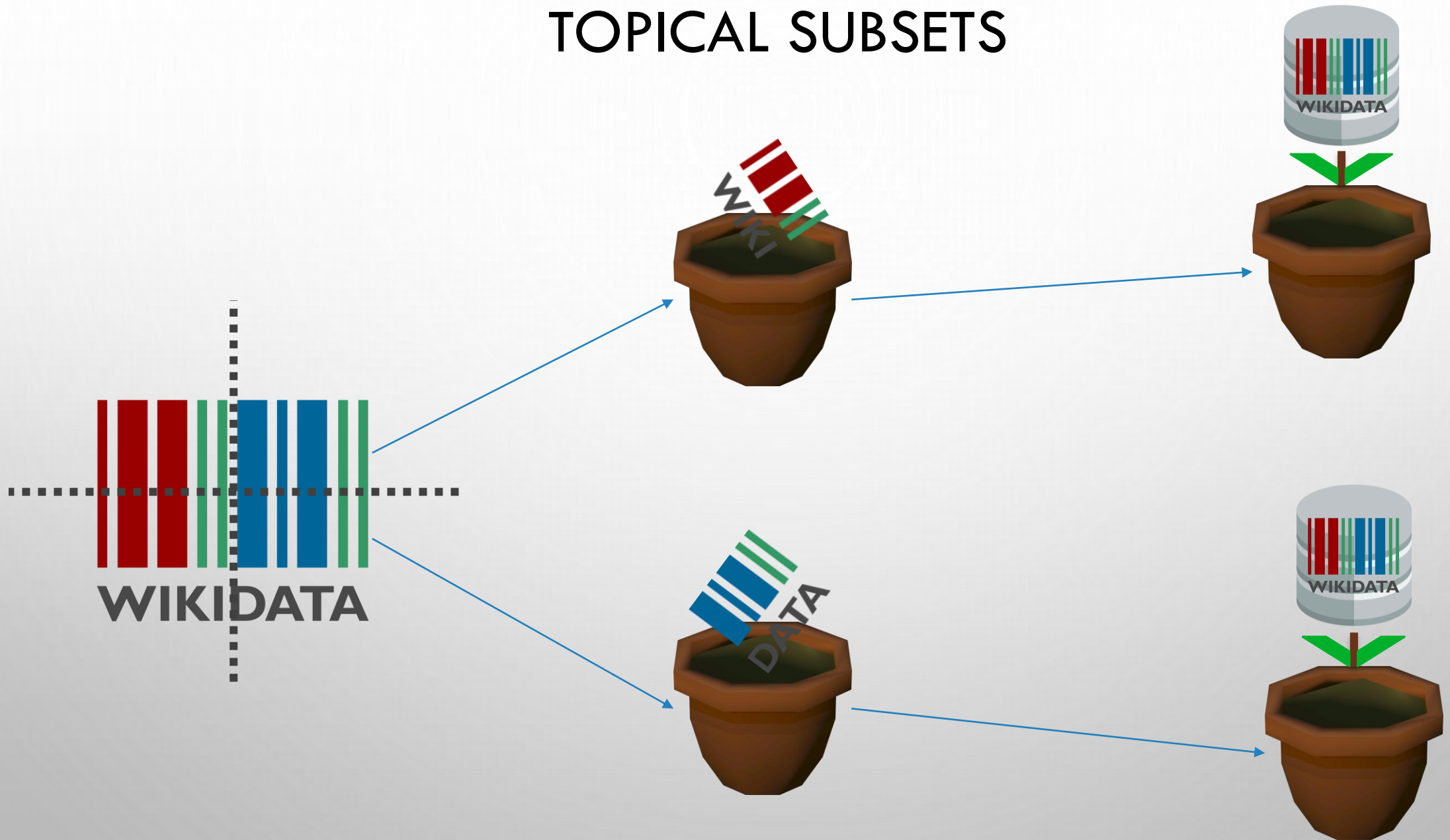


Reducing the Overall Costs



Reproducible Experiments

TOPICAL SUBSETS



SUBSETTING TOOLS

Configuration phase (Filtering): Defining the subset

- Which Items should be extracted?
- Which statements?
- Which metadata (references, qualifiers, labels, ...)?
- Filtering item-based or fact-based?
- Flexibility

Extraction phase: Cutting the defined subset from the main KG

- Be as fast as possible
- Extract accurately (be sure that the output has got what it should got)

DEFINING SUBSET via SHEX

```
PREFIX : <http://example.com/>  
PREFIX wd: <http://www.wikidata.org/entity/>  
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
```

```
start=@:lipids
```

```
:lipids {  
  wdt:P2063 .+;  
  wdt:P234 .+;  
  wdt:P235 .+;  
  wdt:P703 @:taxon +;  
}
```

```
:taxon {  
  wdt:P31 [wd:Q16521];  
}
```

A Simple Lipids subset

https://github.com/seyedahbr/biohackathon2021/tree/main/use_cases/lipidmaps

DEFINING SUBSET via SHEX

```
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>

<#chemical_compound> EXTRA wdt:P31 {
  wdt:P31 [ wd:Q11173 ] | wdt:P279 @<#chemical_compound> + ;
}

<#disease> EXTRA wdt:P31 {
  wdt:P31 [ wd:Q12136 ] | wdt:P279 @<#disease> + ;
}

<#gene> EXTRA wdt:P31 {
  wdt:P31 [ wd:Q7187 ] | wdt:P279 @<#gene> + ;
}

<#protein> EXTRA wdt:P31 {
  wdt:P31 [ wd:Q8054 ] | wdt:P279 @<#protein> + ;
}
```

**Part of GeneWiki subset definition
with considering sub-classes using
recursive**

<https://github.com/seyedahbr/Wikidata-Reference-Statistics/blob/main/Schema%20schemata/genewiki.shex>

REFERENCES

- [1] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, S. Auer, Quality Assessment For Linked Data: A Survey, *Semantic Web*. 7 (2016) 63–93.
- [2] Beghaeiraveri, S. A. H., Gray, A. J., & Mcneill, F. J. (2021, October). Reference Statistics In Wikidata Topical Subsets. In 2nd Wikidata Workshop.