CCKS2016 Tutorial on Testing and Assessing the Quality of Knowledge Graphs

Presenters

Jeff Z. Pan[1], Tong Ruan[2]

[1]University of Aberdeen

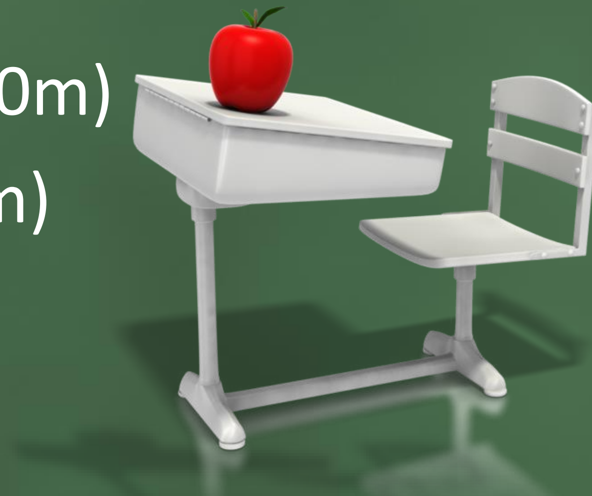
[2]Eest China University of Science and Technology

Agenda

PART I LINKED DATA & KNOWLEDGE GRAPH

2:00 pm – 3:15 pm	**Part 1: Testing the Schema of Knowledge Graphs (Jeff)**
3:45 pm – 5:00 pm	**Part 2: Assessing the Data of Knowledge Graphs (Tong)**

PART I
TESTING AND ASSESSING THE SCHEMA OF KNOWLEDGE GRAPHS



- A Brief History of Knowledge Graphs (20m)
- Test Driven Ontology Construction (10m)
- An Exercise on Competency Questions (30m)
- Generating Authoring Tests from Competency Questions (30m)

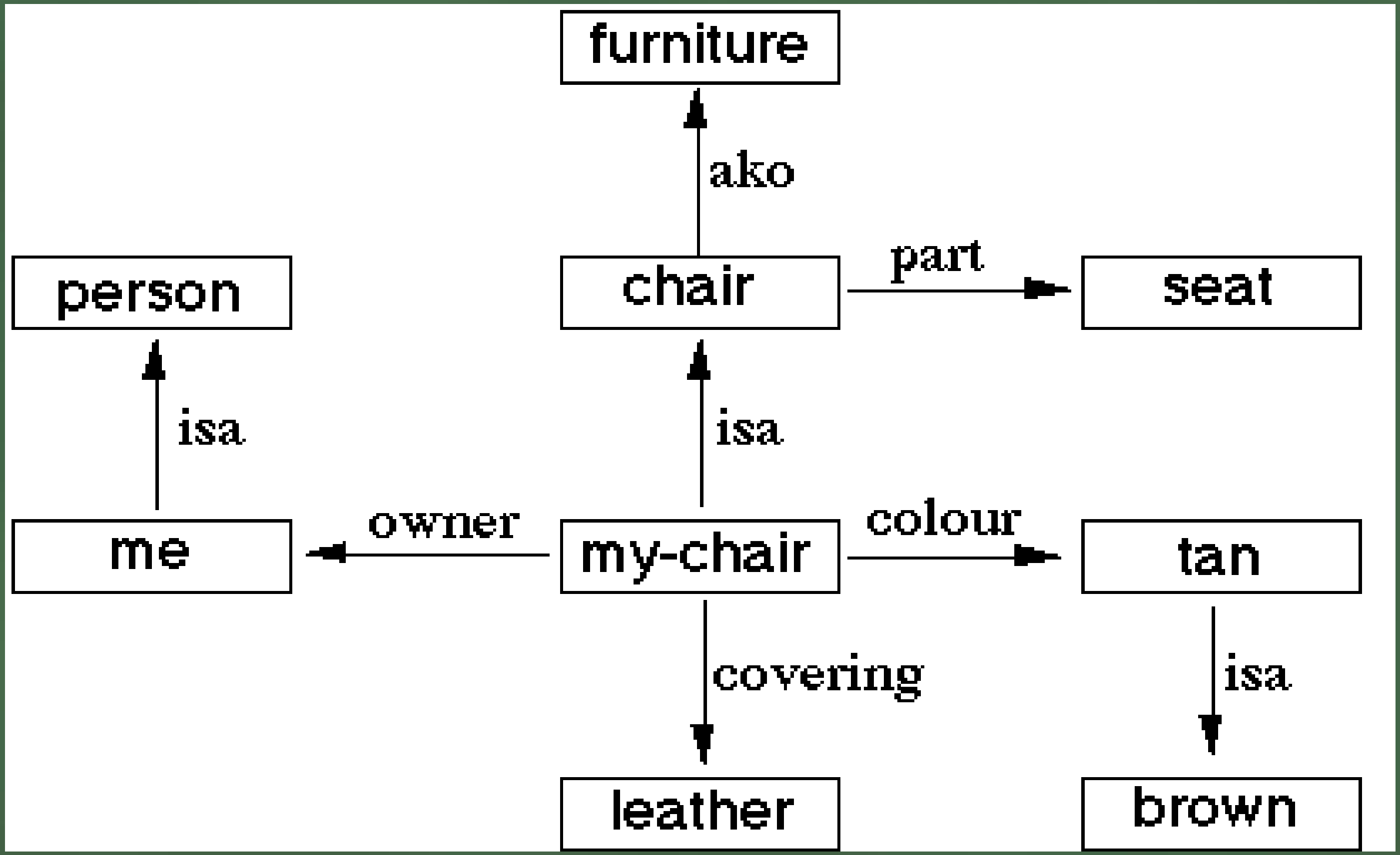
Let's not reinvent the wheel …

A QUICK OVERVIEW OF KNOWLEDGE GRAPHS

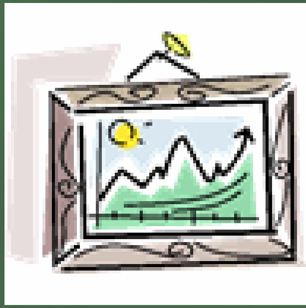
Semantic Nets

- Proposed by Quillian (1968) to analysis the meaning of words in sentences
  - Later applied to KR
- Basic notations:
  - Nodes: to represent objects, concepts, or situations
  - Arcs: to represent relationships

An Example Semantic Net



Frames



- A frame [Minsky 1981] represents knowledge
  - About a narrow subject
  - That has default (*) knowledge
- Example:

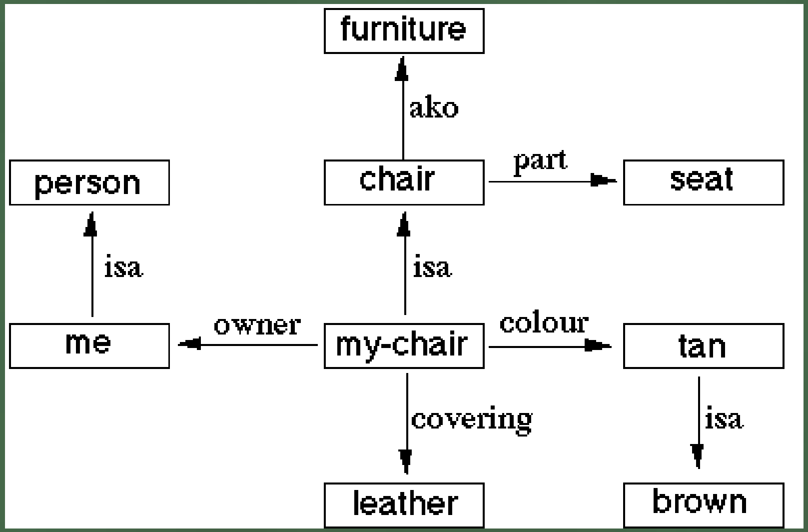
```
Mammal
  subclass: Animal
  warm_blooded: yes
Elephant
  subclass: Mammal
  * colour: grey
  * size:  large
Cari:
 instance: Elephant
 size: small
```

Example: Procedures in Frames

```
age:
   range      1..1500
   if_needed  ask("What  is the age of ", name of this person)
   if_added   if new value > 1500 then
                    print("Your ", this slot, " is too high!");
```

Resource Description Framework

- Modern version of semantic network
  - Basic building block: Subject-property-value
  - Many syntax, including N3 and JSON-LD



[my-chair colour tan .]

[my-chair rdf:type chair .]

[chair rdfs:subClassOf furniture .]

…

Resource Description Framework

- Limited schema support: not expressive enough to cover primary and foreign keys

Student ID	Name	take-course
p001	John	cs3019
p002	Tom	cs3023

– [csd:p001 rdf:type csd:Student .]
– [csd:p002 rdf:type csd:Student .]
– [csd:p001 csd:name “John” .]
– [csd:p002 csd:name “Tom” .]
– [csd:p001 csd:take-course csd:cs3019 .]
– [csd:p002 csd:take-course csd:cs3023 .]

KL-ONE

- Use FOL-based formal semantics, but limit the expressive power
- Formalising Semantic Nets
  - Formal semantics for built-in relationships
    - Sub-Class-Of
    - Instance-of
- Provide constructors to define concepts
- the use of an automatic deductive classifier
- Later, this kind of KR language is called Description Logics (underpinnings of W3C OWL standard)

But where the tests come from?

TEST DRIVEN ONTOLOGY CONSTRUCTION

Is it as simple as it seems?

AN EXERCISE ON COMPETENCY QUESTIONS

Online Exercise

- URL: http://homepages.abdn.ac.uk/jeff.z.pan/pages/link/WhatIf/

http://homepages.abdn.ac.uk/jeff.z.pan/pages/link/WhatIf/

Voluntary Research Survey

Study Information

This study is administered by the Department of Computing Science at the University of Aberdeen.

In this study, we will ask you to rate the feedback from a new ontology authoring tool.

First, we will collect some demographic information about you (age, gender, occupational background and experience with ontology creation).

We will then show you an ontology created by an ontology engineer, and ask you for your opinions on the feedback the ontology authoring tool is giving to them.

Your participation is entirely voluntary and you may withdraw from this study at any time. Your data will be held securely, and no personally idenifiable data will be published.

If you have any questions about this study, please email k.vdeemter[at]abdn.ac.uk.

Consent Form

Please read the statements below and check the final box to confirm you have read and understood the statements and upon doing so agree to participate in the project.

- I confirm that this questionnaire has been explained to me. I have had the opportunity to ask questions about the project and have had these answered satisfactorily.
- I consent to the material I contribute being used to generate insights for the Competency Questions Survey.
- I understand that my participation in this research is voluntary and that I may withdraw from the study at any time.
- I consent to allow the fully anonymised data to be used for future publications and other scholarly means of disseminating the findings from the research project.
- I understand that the information/data acquired will be securely stored by researchers, but that appropriately anonymised data may in future be made available to others for research purposes only.
- I understand that I can request any of the data collected from/by me to be deleted.

Please enter your email address:

Email address

☐ I confirm I have read and understood the above statements, and give my consent to participate in this study.

Begin

http://homepages.abdn.ac.uk/jeff.z.pan/pages/link/WhatIf/

Voluntary Research Survey

About You

Q1. Please indicate your age:

18-25 | 26-40 | 41-65 | over 65 | rather not say

Q2. Please indicate your gender

Male | Female | rather not say

Q3. What is your occupational background?

occupation

Q4. What is your approximate experience creating and editing Ontologies?

- Novice (just started)
- Beginner (up to 1 year)
- Intermediate (1-2 years)
- Expert (over 2 years)

Next

Voluntary Research Survey

Instructions

Costabucks is a hot drinks company. They are creating a robot that can answer questions from customers about the hot drinks that they sell. The robot's programmers have to tell the robot some facts about hot drinks so that it understands enough to answer the questions. To do this, the robot's designers give the robot axioms about the world.

Once all of the Customer Questions (CQs) can be answered by the robot, its knowledge of the coffee menu is considered complete, and it can be used in the shop.

The programmers have made a first attempt at the set of axioms, shown below:

1. $hasContent \cdot hasContent \sqsubseteq hasContent$
2. $Drink \equiv CoffeeDrink \sqcup TeaDrink$
3. $Coffee \sqsubseteq \exists hasContent.Caffeine$
4. $CoffeeDrink \equiv \exists\ hasContent.Coffee$
5. $TeaDrink \equiv \exists\ hasContent.Tea$
6. $CoffeeDrink \sqcap TeaDrink \sqsubseteq \bot$
7. $Cappuccino \sqsubseteq Drink \sqcap \exists\ hasContent.SteamedMilk \sqcap \exists\ hasContent.Coffee$
8. $Americano \sqsubseteq CoffeeDrink$

The programmers are using a special **programming tool** which allows them to add possible customer questions to its interface, and the tool can inform them when the questions are able to be answered by the current set of axioms.

To do this, the tool breaks down the questions into several smaller **authoring tests**, which all must be passed in order for the question to be judged as answerable. The authoring tests are automatically generated by the tool, based on what the customer question is.

Example CQ: Which Coffees contain milk?

Authoring Tests:

1. *CoffeeDrink* should occur in the ontology ✔
2. *Milk* should occur in the ontology ✔
3. *hasContent* should occur in the ontology ✔
4. $CoffeeDrink \sqcap \exists hasContent.Milk$ should be satisfiable in the ontology ✔
5. $CoffeeDrink \sqcap \neg(\exists hasContent.Milk)$ should be satisfiable in the ontology ✔

In the example above, the tool has judged this CQ as **answerable**, as the authoring tests have passed (green with a tick ✔).

When you click **"begin"**, you will see 7 questions in the same format. We will show you each question in turn, and its associated authoring tests, and whether they pass or fail. We will then ask you whether you agree if the CQs can be meaningfully answered.

This survey is about whether the *Authoring Tests* are appropriate, rather than the validity of the questions or the axioms.

Begin

Let’s not reinvent the wheel …

GENERATING AUTHORING TESTS FROM COMPETENCY QUESTIONS

Competency Question as Requirements

- A typical CQ: Which pizza has some cheese topping?

- Questions that people expect the constructed ontologies to answer
- Useful for novice users:
  - in natural languages
  - about domain knowledge
  - requires little understanding of ontology technologies

CQs in Ontology Authoring

- A typical CQ: Which pizza has some cheese topping?

- Existing work focused on answering CQs directly

- Applying CQs in ontology authoring
  - Proposed by M. Uschold, M. Gruninger, et al.
  - NeON methodology
  - Visualisation based on goal modelling
  - Formalising CQs into SPARQL queries
  - Formalising CQs into DL queries
  - Natural language CQ understanding

CQs in Ontology Authoring

- A typical CQ: Which pizza has some cheese topping?

- Existing work focused on answering CQs directly
  - But is the answer meaningful?

- The ability to **answer CQs meaningfully** can be regarded as a functional requirement of the ontology

- What if the answer is an empty set

- Possible scenarios
  - Pizza does not exist
  - Cheese topping does not exist
  - Pizzas are not allowed to have cheese topping
  - The ontology has not been populated with any cheesy pizza yet
  - …

Q1: what forms of requirements should we consider

Q1’: How are CQs formulated?

Key Idea 1: Identification of CQ Patterns

- A typical CQ: Which pizza has some cheese topping?

CE1 OPE CE2

Element: Class expression CE2

Feature: Binary predicate

Element: Object property expressions OPE

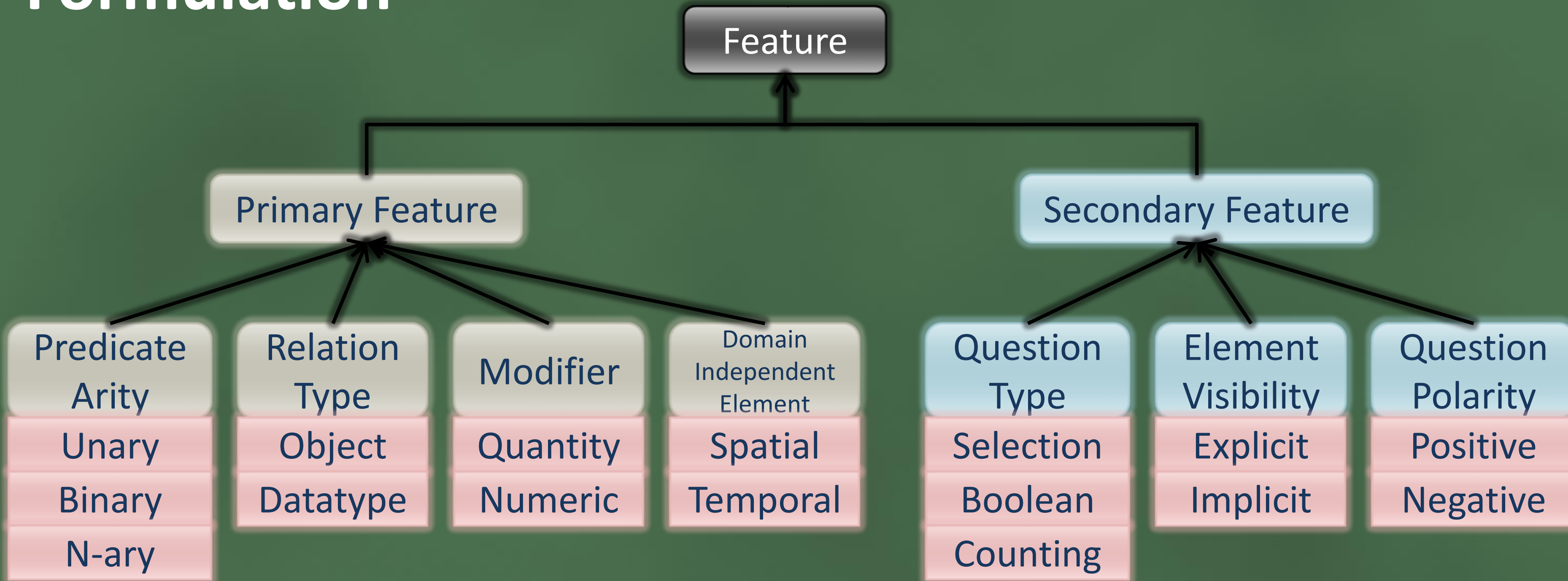
Element: Class expression CE1

Feature: Type of question

- **Hypothesis**: CQs usually have clear syntactic patterns

- Features and elements can be extracted

Result 1: A Feature-based Framework for CQ Formulation



- Based on CQs collected from the Software Ontology Project (75 CQs) and Manchester OWL Workshops (70 CQs)
- Primary features -> CQ Archetypes
- Secondary features -> CQ Subtypes

Result 2: Archetypes of CQ Patterns

Table 1. CQ Archetypes (PA = Predicate Arity, RT = Relation Type, M = Modifier, DE = Domain-independent Element; obj. = object property relation, data. = datatype property relation, num. = numeric modifier, quan. = quantitative modifier, tem. = temporal element, spa. = spatial element; CE = class expression, OPE = object property expression, DP = datatype property, I = individual, NM = numeric modifier, PE = property expression, QM = quantity modifier)

ID	Pattern	Example	PA	RT	M	DE
1	Which [CE1] [OPE] [CE2]?	Which pizzas contain pork?	2	obj.		
2	How much does [CE] [DP]?	How much does Margherita Pizza weigh?	2	data.		
3	What type of [CE] is [I]?	What type of software (API, Desktop application etc.) is it?	1			
4	Is the [CE1] [CE2]?	Is the software open source development?	2			
5	What [CE] has the [NM] [DP]?	What pizza has the lowest price?	1	data.	num.	
6	What is the [NM] [CE1] to [OPE] [CE2]?	What is the best/fastest/most robust software to read/edit this data?	[illegible]	both	num.	
7	Where do I [OPE] [CE]?	Where do I get updates?	2	obj.		spa.
8	Which are [CE]?	Which are gluten free bases?	1			
9	When did/was [CE] [PE]?	When was the 1.0 version released?	2	data.		tem.
10	What [CE1] do I need to [OPE] [CE2]?	What hardware do I need to run this software?	3	obj.		
11	Which [CE1] [OPE] [QM] [CE2]?	Which pizza has the most toppings?	2	obj.	quan.	
12	Do [CE1] have [QM] values of [DP]?	Do pizzas have different values of size?	2	data.	quan.	

Result 2: Subtypes in An Archetype

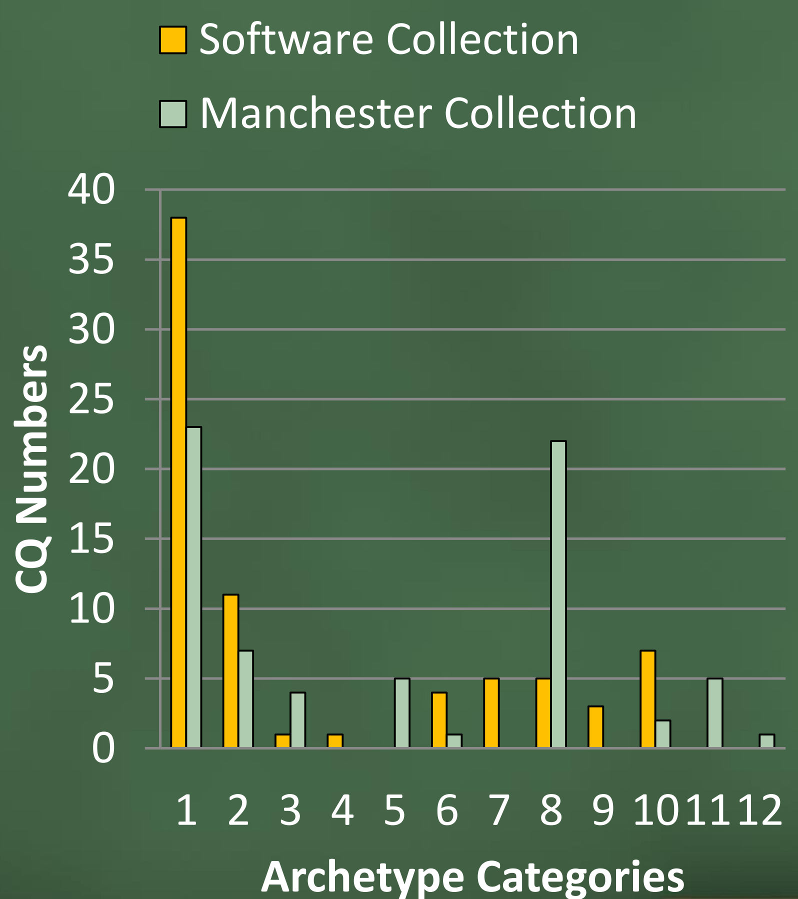
- An example on archetype 1

Table 2. CQ Sub-types of Archetype 1 (QT = Question Type, V = Visibility, QP = Question Polarity, sel. = selection question, bin. = binary question, cout. = counting question, exp. = explicit, imp. = implicit, sub. = subject, pre. = predicate, pos. = positive, neg. = negative)

ID	Pattern	Example	QT	V	QP
1a	Which [CE1] [OPE] [CE2]?	What software can read a .cel file?	sel.	exp.	pos.
1b	Find [CE1] with [CE2].	Find pizzas with peppers and olives.	sel.	imp. pre.	pos.
1c	How many [CE1] [OPE] [CE2]?	How many pizzas in the menu contains meat?	cout.	exp.	pos.
1d	Does [CE1] [OPE] [CE2]?	Does this software provide XML editing	bin.	exp.	pos.
1e	Be there [CE1] with [CE2]?	Are there any pizzas with chocolate?	bin.	imp. pre.	pos.
1f	Who [OPE] [CE]?	Who owns the copyright?	sel.	imp. sub.	pos.
1g	Which [CE1] [OPE] no [CE2]?	Which pizza contains no mushroom?	sel.	exp.	neg.

Result 3: Formulation of Real-world CQs

- 6 archetypes are shared by the 2 collections
  - They are also the most populated archetypes
  - Cover **86.2%** of total CQs in our collections
- Coverage on CQs studied by existing works: 55 CQs in total
  - All except 1 covered by our framework
  - Archetype 1 is still the most populated
  - The only remaining one is *"Why universities are organised into departments?"*



Q2: how to generate authoring tests from requirements

Q2’: How can we automatically test whether a CQ can be meaningfully answered?

Key Idea 2: Presuppositions of CQ

- A typical CQ: Which pizza has some cheese topping?

- A CQ comes with certain *presuppositions*
  - *Some conditions the speakers assume to be met*
- A CQ can be *meaningfully answered* only when its presuppositions are satisfied

- Classes *Pizza*, *CheeseTopping* should occur in the ontology
- Property *has(Topping)* should occur in the ontology
- The ontology should *allow Pizza to have CheeseTopping*
- The ontology should also *allow Pizza to not have CheeseTopping*

CQs in Ontology Authoring

- A typical CQ: Which pizza has some cheese topping?

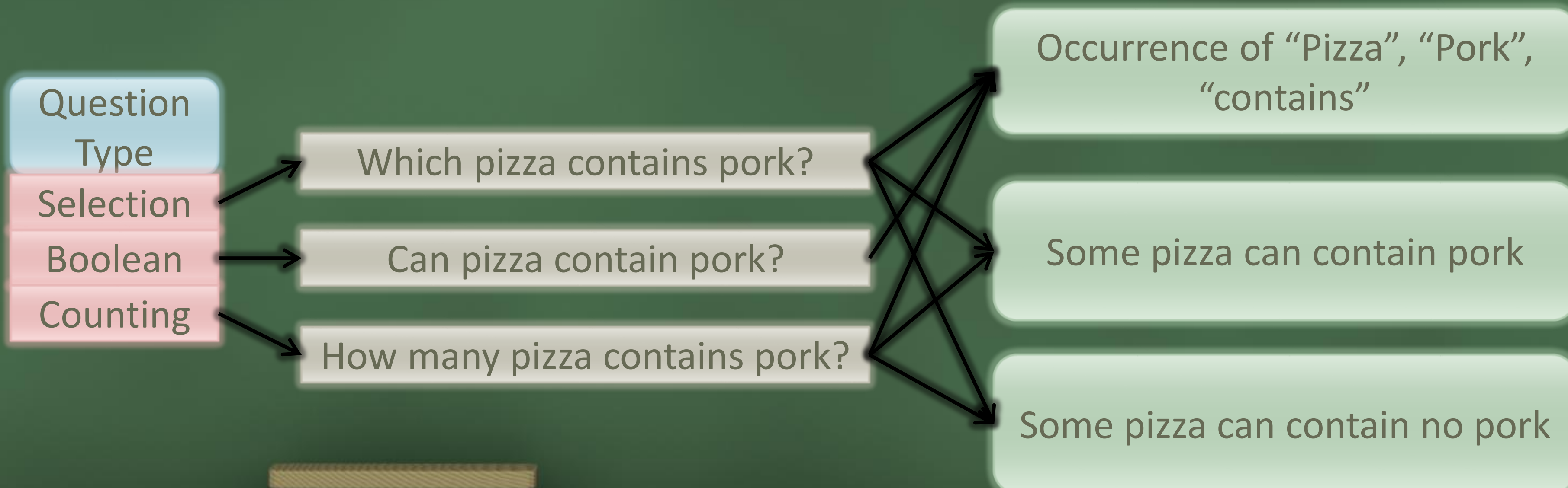
CE1 OPE CE2

- Satisfiability of CQ presuppositions can be verified by **authoring tests** generated based on its features and elements

- Classes *Pizza*, *CheeseTopping* should occur in the ontology
  - [CE1], [CE2] should both ***occur*** in the class vocabulary
- Property *has(Topping)* should occur in the ontology
  - [OPE] should ***occur*** in the property vocabulry
- The ontology should *allow Pizza to have CheeseTopping*
  - should be ***satisfiable***
- The ontology should also *allow Pizza to not have CheeseTopping*
  - should be ***satisfiable***

Result 4: Associate Presuppositions with Features

- Features in a CQ are associated with the presuppositions of the CQ.
  - An example on the ***question type*** feature:



Result 5: Formal Authoring Tests

Table 4. Authoring Tests ($\sqcap$ means conjunction, $\neg$ means negation, $\exists P.E$ means having P relation to some E, $= nP.E$ ($\geq nP.E$, $\leq nP.E$) means having P relation(s) to exactly (at least, at most) n E(s), $\forall P.E$ means having P relation (if any) to only E, $\top$ means everything)

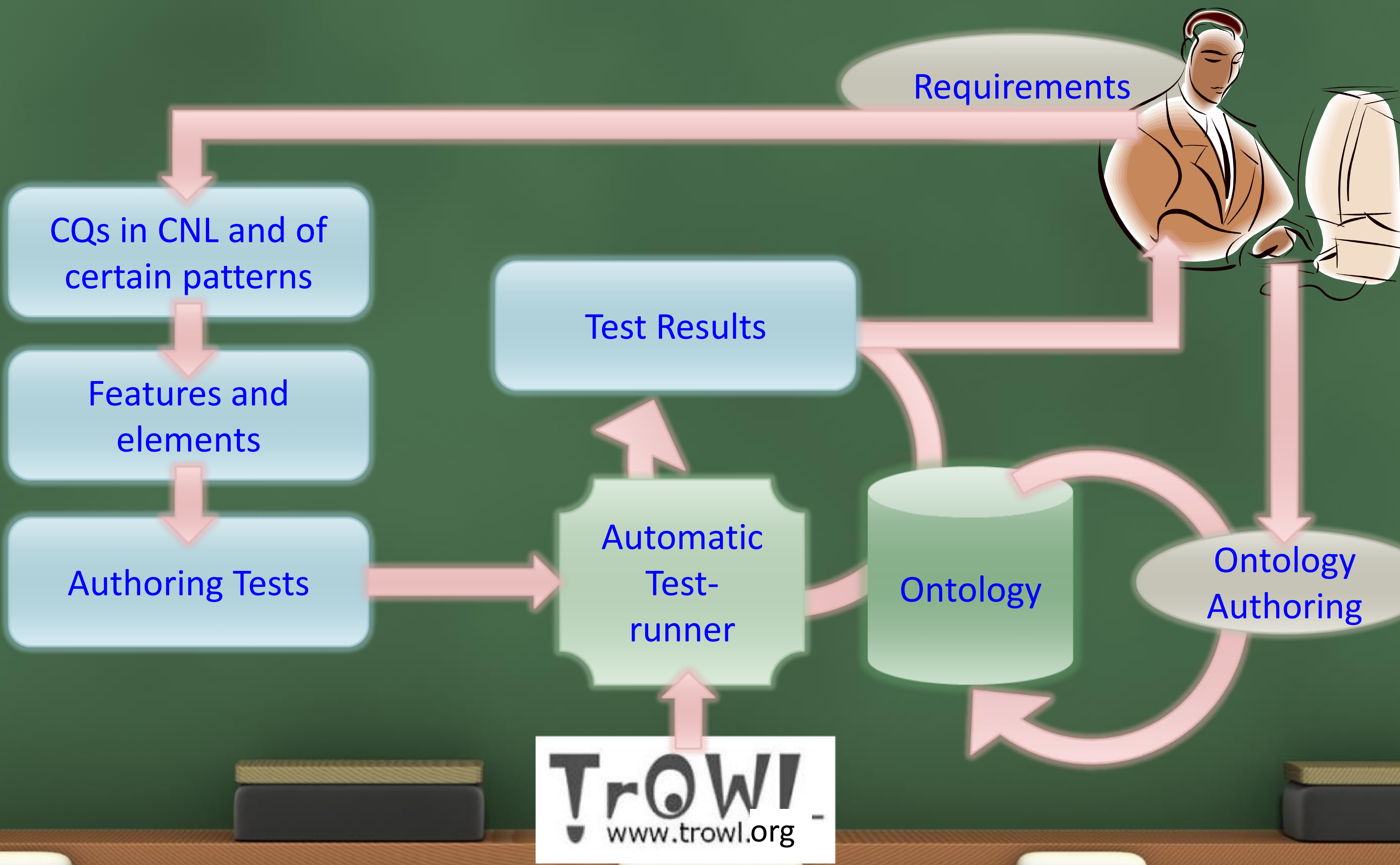
AT	Parameter	Realisation
Occurrence	[E]	E in ontology vocabulary
Class Satisfiability	[CE]	CE is satisfiable
Relation Satisfiability	[CE1] [P] [E2]	$CE1 \sqcap \exists P.E2$ is satisfiable, $CE1 \sqcap \neg\exists P.E2$ is satisfiable
Meta-Instance	[E1] [E2]	$E1$ has `type` $E2$
Cardinality Satisfiability	[CE1] [n] [P] [E2]	$CE1 \sqcap = nP.E2$ is satisfiable, $CE1 \sqcap \neg = nP.E2$ is satisfiable
Multiple Cardinality (on concrete quantity modifier)	[CE1] [P] [E2]	$\forall n \geq 0, CE1 \sqcap \neg = nP.E2$ is satisfiable
Comparative Cardinality (on quantity modifier)	[CE1] [P] [E2]	$\exists n \geq 0, CE1 \sqcap \neg \leq n\ P.E2$ and $CE1 \sqcap \geq n{+}1\ P.E2$ are satisfiable, $\exists m \geq 0, CE1 \sqcap \neg \leq m\ P.E2$ and $CE2 \sqcap \geq m\ P.E2$ are satisfiable
Multiple Value (on superlative numeric modifier)	[CE1] [P]	$\forall D \subseteq range(P), CE1 \sqcap \neg\exists P.D$ is satisfiable
Range	[P] [E]	$\top \sqsubseteq \forall P.E$

- All testings can be automated

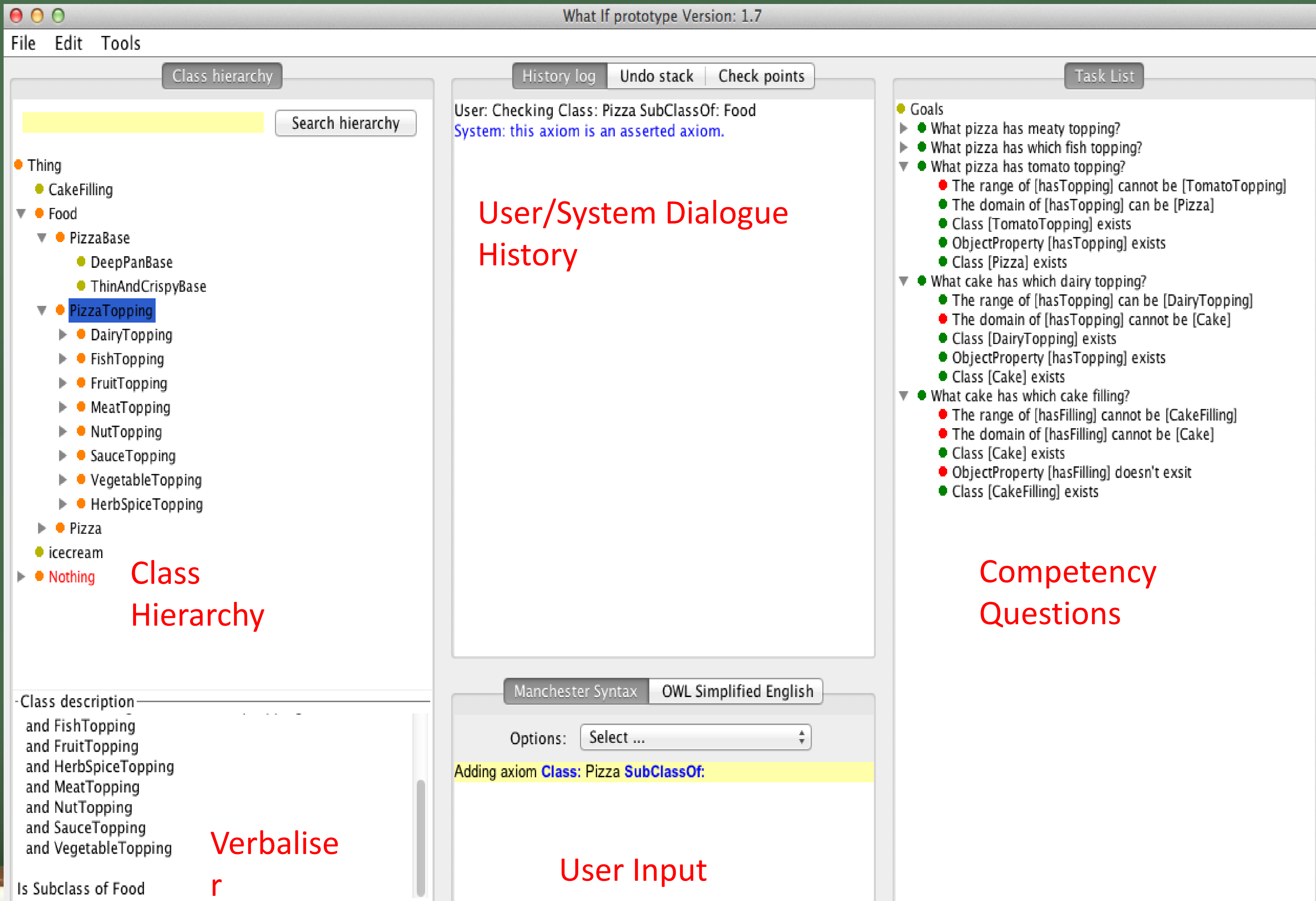
Reasoning Implications (1)

- Services required:
  - Concept satisfiability checking
  - Subsumption checking
- Expressive power needed: In general ALCQ
  - Pay as you go for EL ontologies
    - All the cardinality related presuppositions are satisfied
    - Testing the satisfiability of CE1 $\neg\exists$P.E2 can be achieved by testing the non-entailment of CE1 $\exists$P.E2
- Possible optimisations:
  - Certain tests may be subsumed by others

A Competency Question-driven Ontology Authoring Pipeline

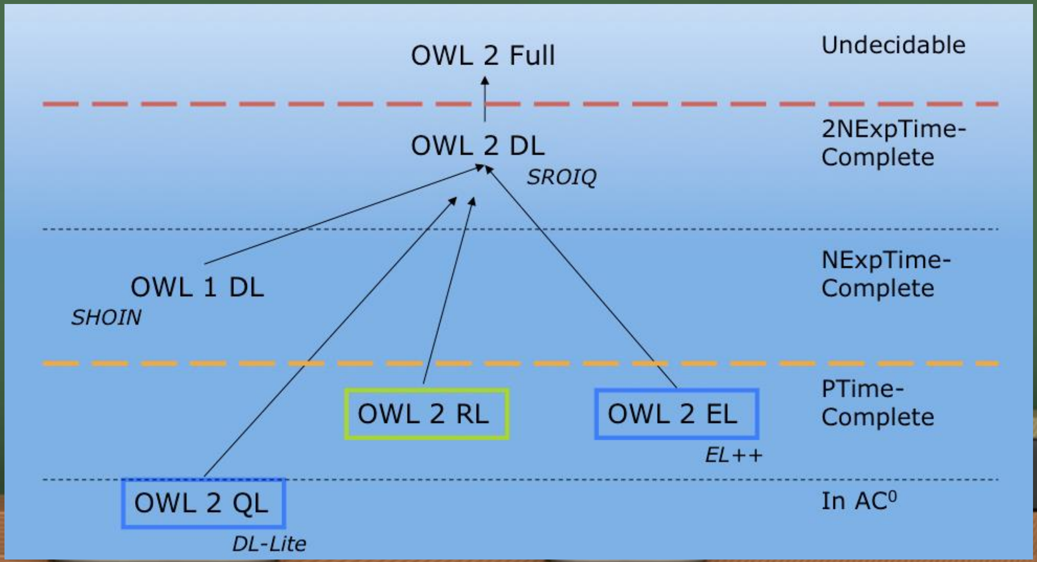


WhatIf Gadget



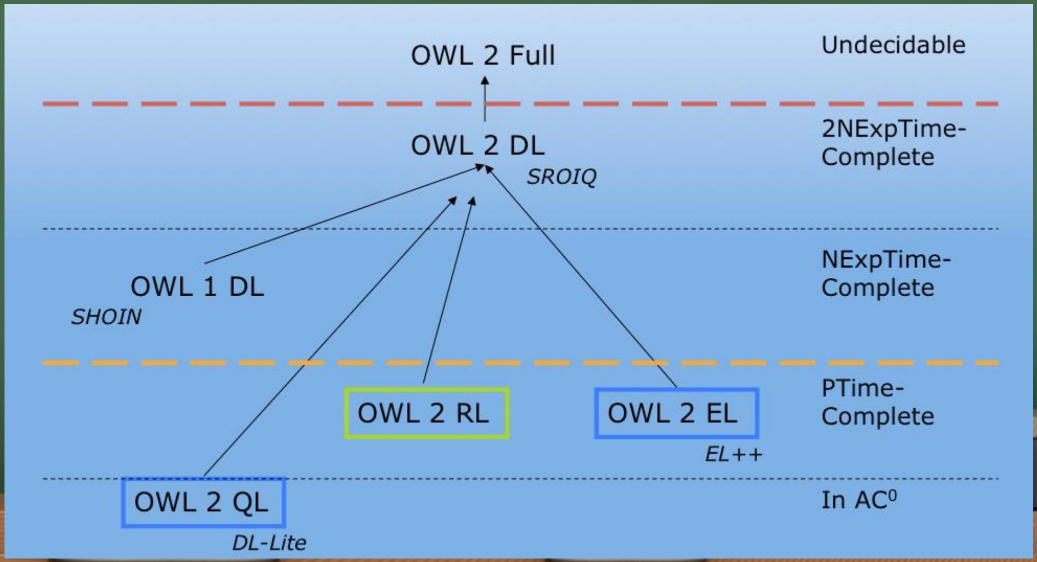
Reasoning Implications (2)

- Responsive reasoning services required:
  - tractable ontology languages
  - faithful approximate reasoning [PanThomas-AAAI2007, RenPanZhao-AAAI2010, Fokoue-etal-AAAI2012, PanRenZhao-AIJ2016]



Reasoning Implications (3)

- Incremental reasoning services required:
  - book-keeping / non-book-keeping
  - forward chaining / backward chairing
  - A Combined approach to Incremental Reasoning for EL Ontologies [Ren-et.al-RR2016]



Summary and Outlook

- How are real-world CQs formulated?
  - Most real-world CQs we have collected can be described with a feature-based framework
- How can we automatically test whether a CQ can be meaningfully answered?
  - Presuppositions of CQs play an important role
  - CQs and authoring tests also helpful for collaborative authoring
- Demand responsive and incremental reasoning
- Future work
  - Implement the pipeline and further evaluations
  - Further extend the CQ description framework and presuppositions
  - New reasoning algorithms and optimisations

PART II
ASSESSING THE DATA OF KNOWLEDGE GRAPHS

- An Introduction to Data Quality Problems
- Models and Metrics of Data Quality Assessment
- Tools and Standards of Data Quality Assessment
- Challenges in Data Quality Assessment
- Some Recent Results

Examples of Data quality problems (1)

- Data inconsistency in EMR

	empi	name	sex	jzlsh	hospital_id	outpatient_id	outdept_name
1	E20140701E330504	李[redacted]	女	01217055	01217055	0060463367	肿瘤科
2	E20140701E330504	李[redacted]	女	01304638	01304638	0060463367	肿瘤科
3	E20140701E330504	李[redacted]	女	01210903	01210903	0060463367	肿瘤科
4	E20140701E330504	李[redacted]	女	01309387	01309387	0060463367	肿瘤科
5	E20140701E330504	李[redacted]	男	01202653	01202653	0060463367	肿瘤科
6	E20140701E330504	李[redacted]	女	01317361	01317361	0060463367	肿瘤科

-

disease_class	disease_code	disease_name
ICD10	E14.900	2型糖尿病
ICD10	E14.901	2型糖尿病 糖尿病肾病
ICD10	E14.203+	2型糖尿病 糖尿病肾病
ICD10	E14.901	2型糖尿病 糖尿病肾病
ICD10	E14.901	2型糖尿病 糖尿病肾病
ICD10	E11.901	2型糖尿病 糖尿病肾病

Examples of Data quality problems (2)

- Inconsistency in data representation in EMR

E20150615E4333149	01514819	男	64Y
E20150618E5342993	00706619	女	76岁
E20150618E5342993	00708622	女	76岁
E20150618E5342993	00710343	女	76Y
E20150618E5342993	00800770	女	76Y

籍贯
上海
上海市
上海市
上海
上海
上海

E20150615E4349749	00911245	男
E20150615E4349749	00912309	男
E20150615E4349749	01002471	男性
E20150615E4349749	01003615	男
E20150615E4349749	01015045	男
E20150615E4349749	01100214	男

Examples of Data quality problems (3)

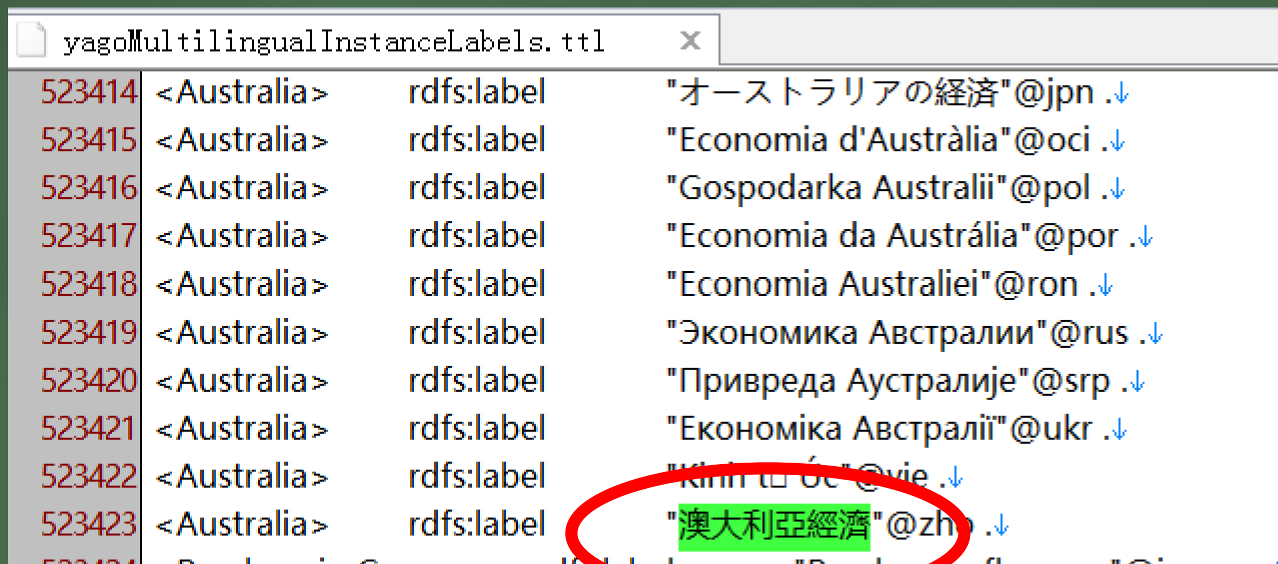
- Wrong property value triple in DBpedia
  - Universe Jones (film), Duration, “USA”

制片人	Dean Devlin
类　型	动作，奇幻，科幻，冒险
主　演	Viveca Lindfors，Kurt Russell，James Spader
片　长	USA:128 min
上映时间	1994年10月28日
MPAA评级	PG-13
原创音乐	David Arnold
视觉特效	Kit West

- Meaningless triple in DBpedia
  - 獾(badger), wasCreatedOnDate, 1758^^xsd:date

Examples of Data quality problems (4)

- Wrong property value triple in YAGO

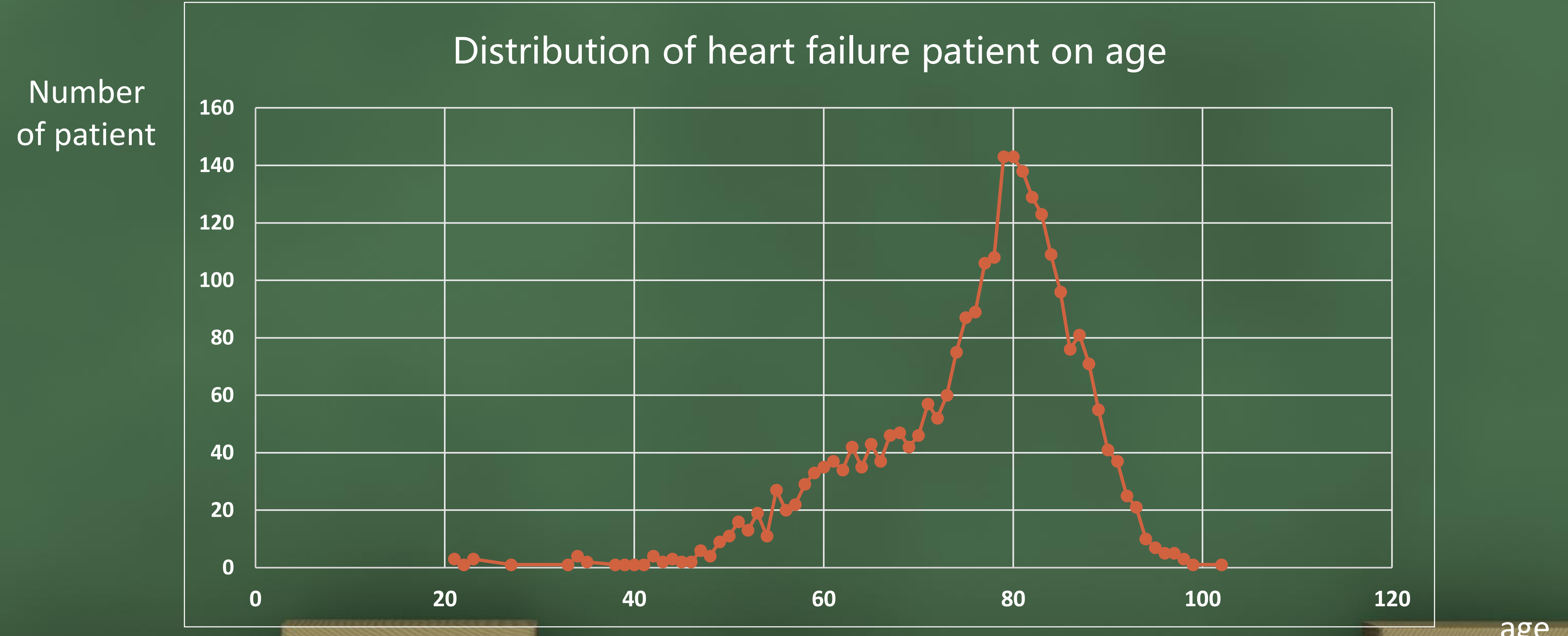

yagoMultilingualInstanceLabels.ttl

523414	<Australia>	rdfs:label	"オーストラリアの経済"@jpn .
523415	<Australia>	rdfs:label	"Economia d'Austràlia"@oci .
523416	<Australia>	rdfs:label	"Gospodarka Australii"@pol .
523417	<Australia>	rdfs:label	"Economia da Austrália"@por .
523418	<Australia>	rdfs:label	"Economia Australiei"@ron .
523419	<Australia>	rdfs:label	"Экономика Австралии"@rus .
523420	<Australia>	rdfs:label	"Привреда Аустралије"@srp .
523421	<Australia>	rdfs:label	"Економіка Австралії"@ukr .
523422	<Australia>	rdfs:label	"Kinh tế Úc"@vie .
523423	<Australia>	rdfs:label	"澳大利亞經濟"@zho .

Economics of Australia

The motivation of data quality assessment (I)

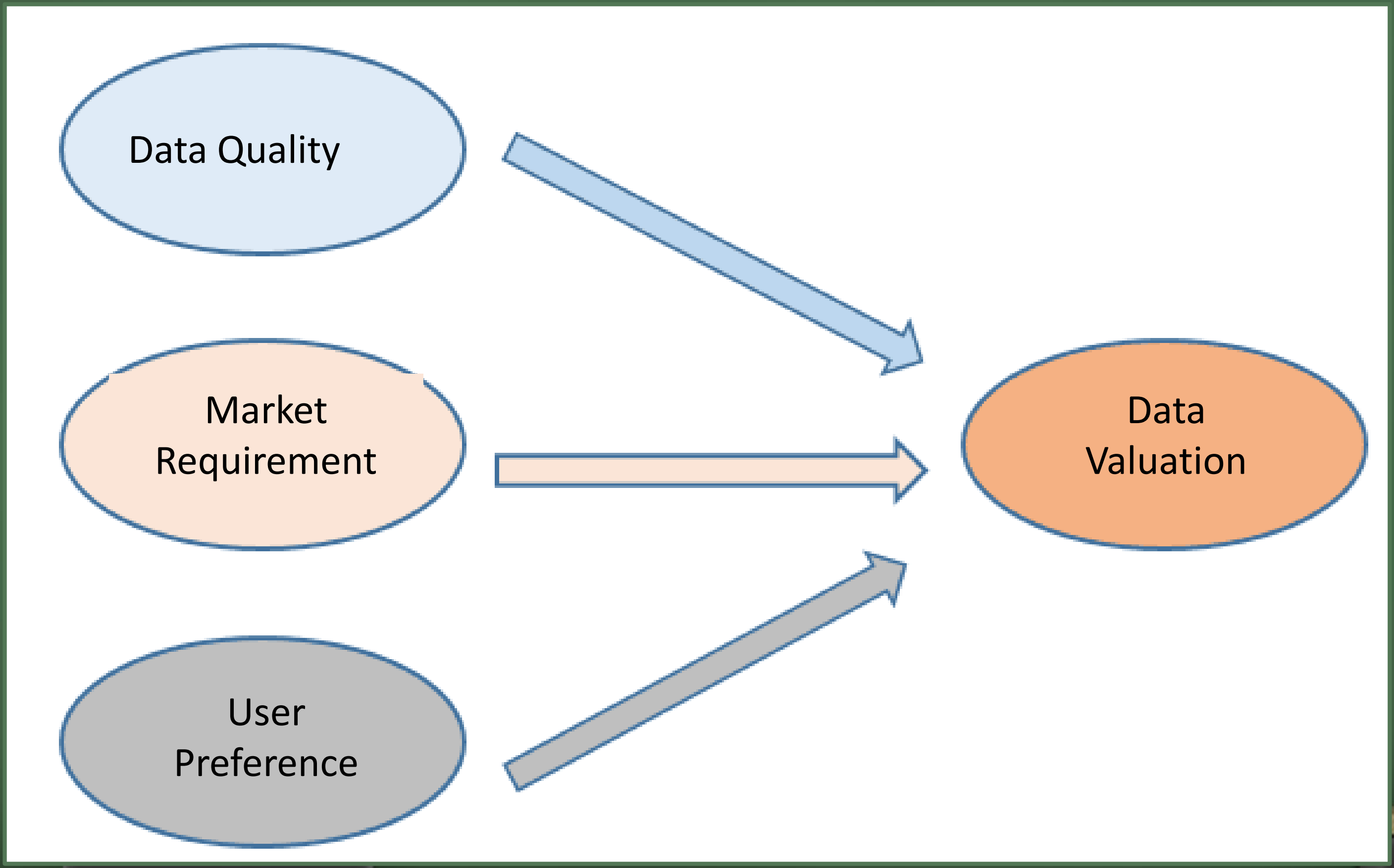
- Data quality assessment is a base of **data usage**
  - Use Electronic Health Record data for risk analysis of heart failure



HUAN-HUAN ZHANG, PENG-CHENG WANG, YING-JIE WANG, TONG RUAN, HAO-FEN WANG. Rish factors of heart failure for patients classification witu extreme learning machine[C]. International conference on machine learning and cybernetics, 2016.

The motivation of data quality assessment (II)

- Data quality assessment is a precursor of **data valuation**

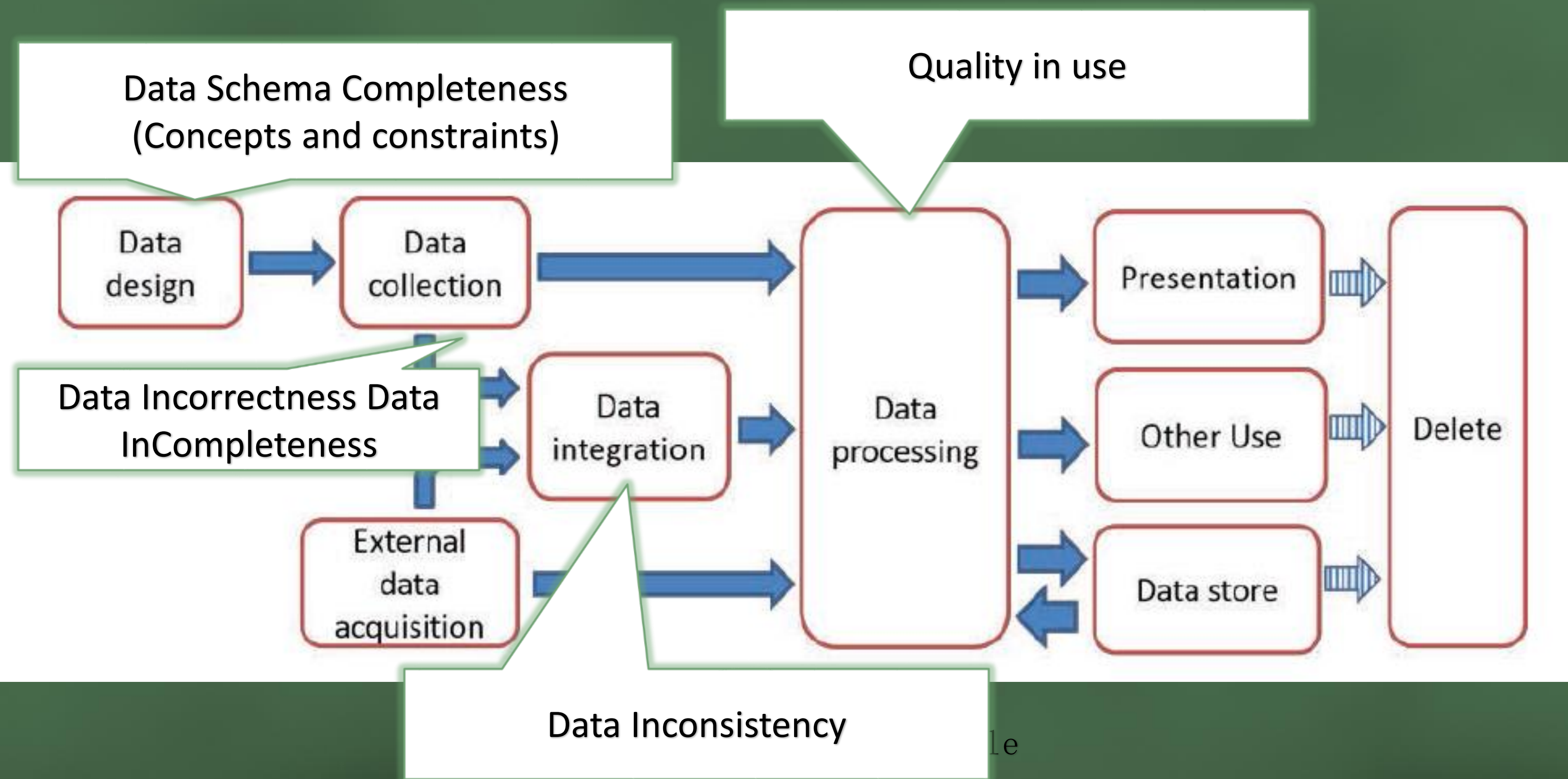


The motivation of data quality assessment (III)

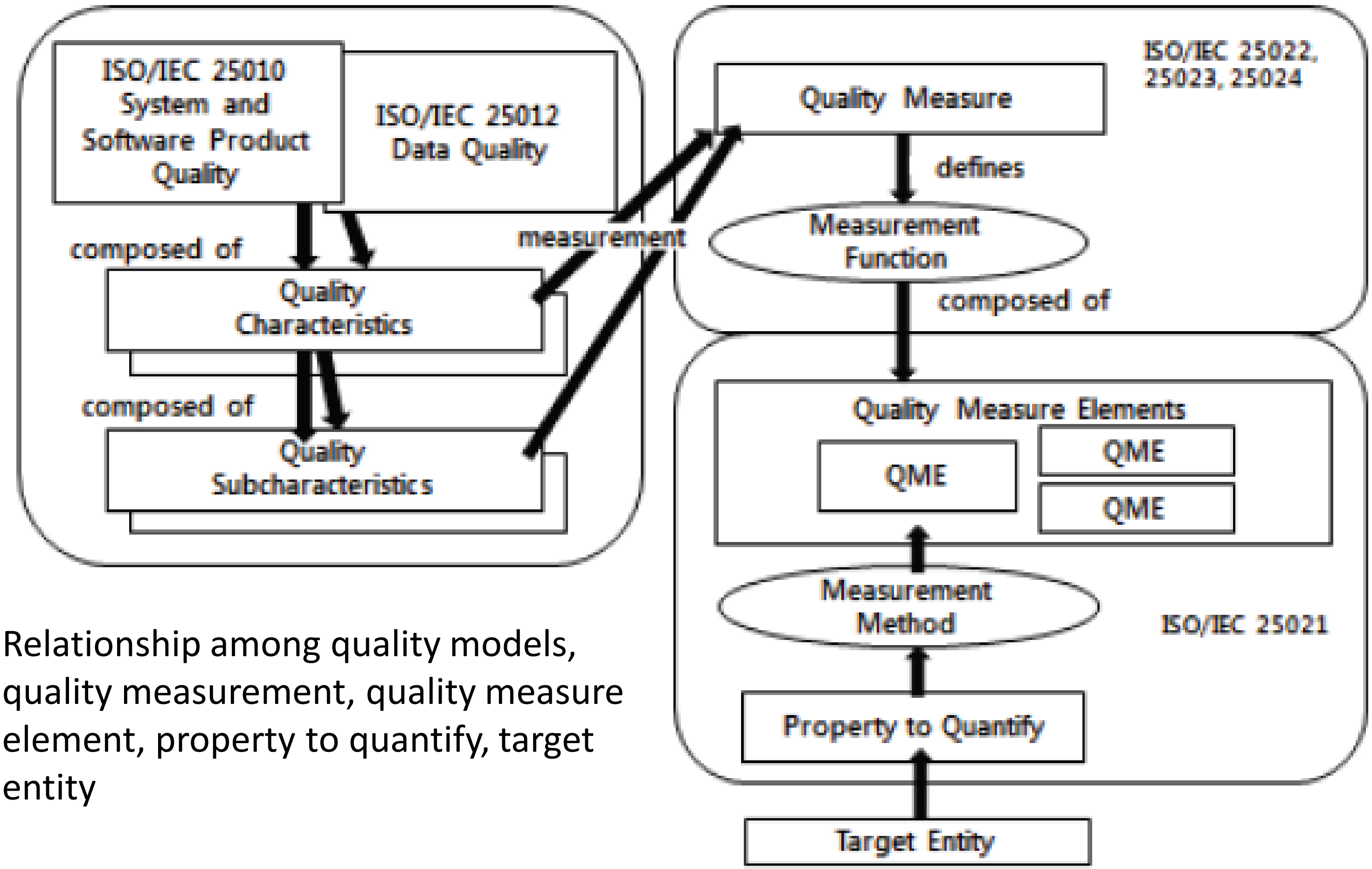
- Data quality assessment is a guidance for **data improvement**

• Redundant data	➡	Remove or fuse redundant data
• Incorrect data detection	➡	Correct data
• Data incompleteness	➡	Data completion
• Different data representation	➡	Data normalization

Quality problems in Data Life Cycle



Relationship among quality models, quality measurement and others.

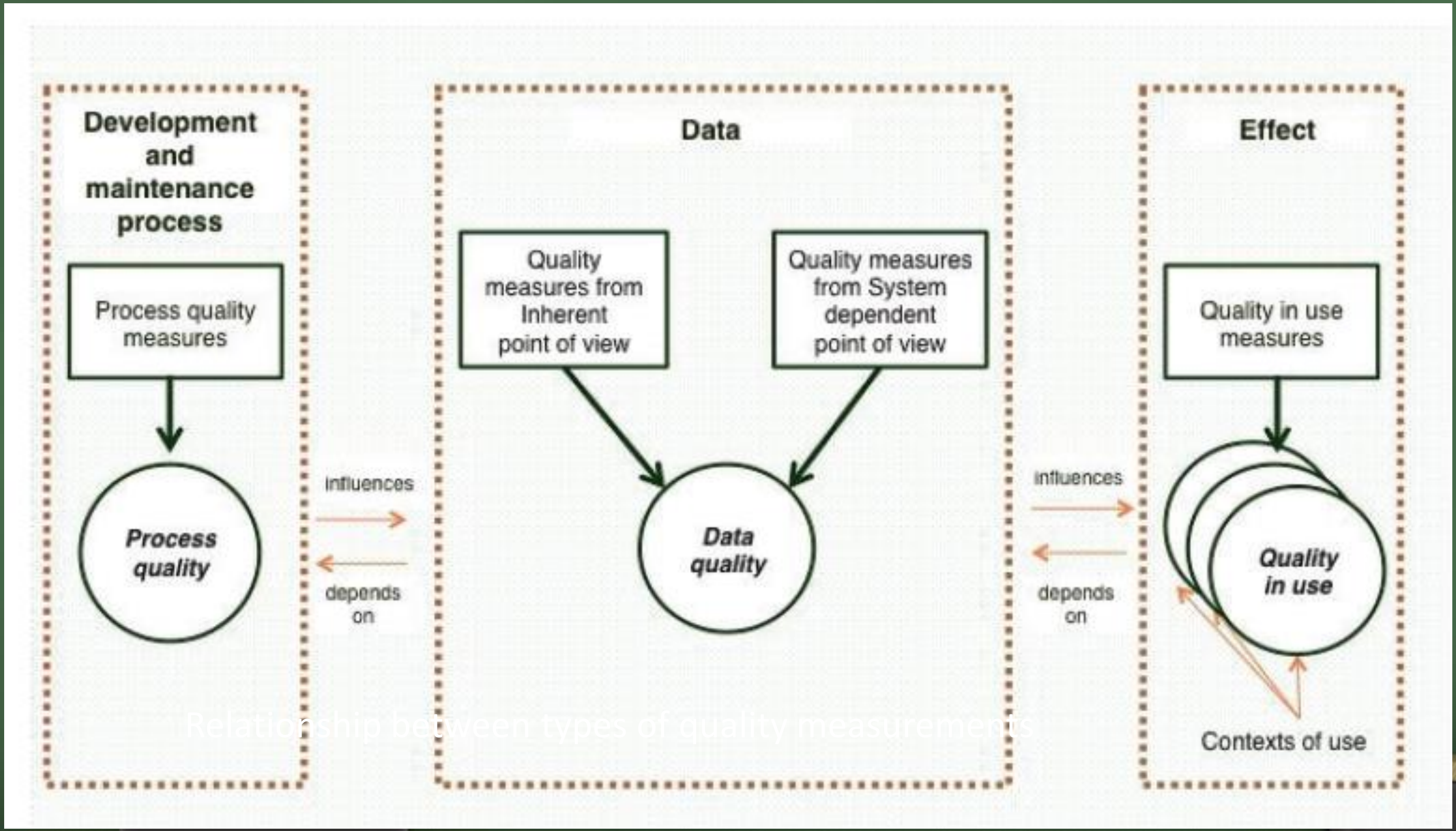


Relationship among quality models, quality measurement, quality measure element, property to quantify, target entity

ISO/IEC 25024:2015(E)

Data quality definition & assessment model

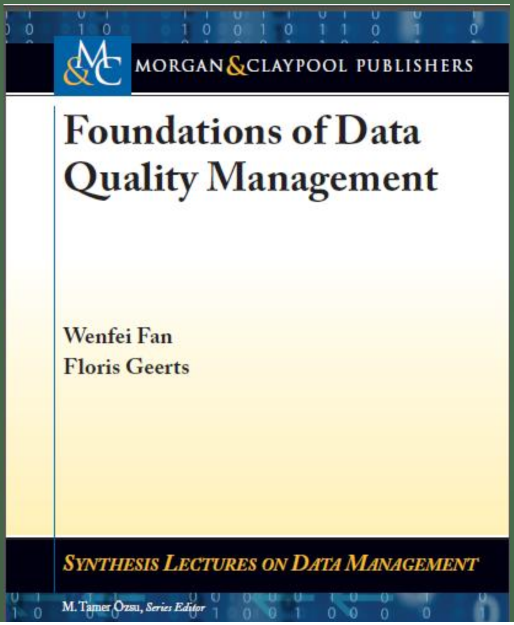
Definition: fitness for In use in special application context



ISO/IEC 25024:2015(E)

Data quality assessment metrics (1)

- Data Consistency, refers to the validity and integrity of data representing real-world entities
- Data Deduplication
- Data Accuracy
- Data Completeness
- Data Currency



	FN	LN	CC	AC	phn	street	city	zip	salary	status
t_1:	Mike	Clark	44	131	null	Mayfield	NYC	EH4 8LE	60k	single
t_2:	Rick	Stark	44	131	3456789	Crichton	NYC	EH4 8LE	96k	married
t_3:	Joe	Brady	01	908	7966899	Mtn Ave	NYC	NJ 07974	90k	married
t_4:	Mary	Smith	01	908	7966899	Mtn Ave	MH	NJ 07974	50k	single
t_5:	Mary	Luth	01	908	7966899	Mtn Ave	MH	NJ 07974	50k	married
t_6:	Mary	Luth	44	131	3456789	Mayfield	EDI	EH4 8LE	80k	married

Data quality assessment metrics (1)

- Data Consistency
- Data Deduplication, aims to identify tuples in one or more relations that refer to the same real-world entity
- Data Accuracy
- Data Completeness
- Data Currency

	FN	LN	CC	AC	phn	street	city	zip	salary	status
t_1:	Mike	Clark	44	131	null	Mayfield	NYC	EH4 8LE	60k	single
t_2:	Rick	Stark	44	131	3456789	Crichton	NYC	EH4 8LE	96k	married
t_3:	Joe	Brady	01	908	7966899	Mtn Ave	NYC	NJ 07974	90k	married
t_4:	Mary	Smith	01	908	7966899	Mtn Ave	MH	NJ 07974	50k	single
t_5:	Mary	Luth	01	908	7966899	Mtn Ave	MH	NJ 07974	50k	married
t_6:	Mary	Luth	44	131	3456789	Mayfield	EDI	EH4 8LE	80k	married

Data quality assessment metrics (1)

- Data Consistency
- Data Deduplication
- Data Accuracy, refers to the closeness of values in a database to the true values of the entities that the data in the database represent
- Data Completeness
- Data Currency

	FN	LN	age	height	status
s_0:	Mike	Clark	14	1.70	single
s_1:	M.	Clark	14	1.69	married
s_2:	Mike	Clark	45	1.60	single

Data quality assessment metrics (1)

- Data Consistency
- Data Deduplication
- Data Accuracy
- Data Completeness, concerns whether our database has complete information to answer our queries
- Data Currency

	FN	LN	CC	AC	phn	street	city	zip	salary	status
t_1:	Mike	Clark	44	131	null	Mayfield	NYC	EH4 8LE	60k	single
t_2:	Rick	Stark	44	131	3456789	Crichton	NYC	EH4 8LE	96k	married
t_3:	Joe	Brady	01	908	7966899	Mtn Ave	NYC	NJ 07974	90k	married
t_4:	Mary	Smith	01	908	7966899	Mtn Ave	MH	NJ 07974	50k	single
t_5:	Mary	Luth	01	908	7966899	Mtn Ave	MH	NJ 07974	50k	married
t_6:	Mary	Luth	44	131	3456789	Mayfield	EDI	EH4 8LE	80k	married

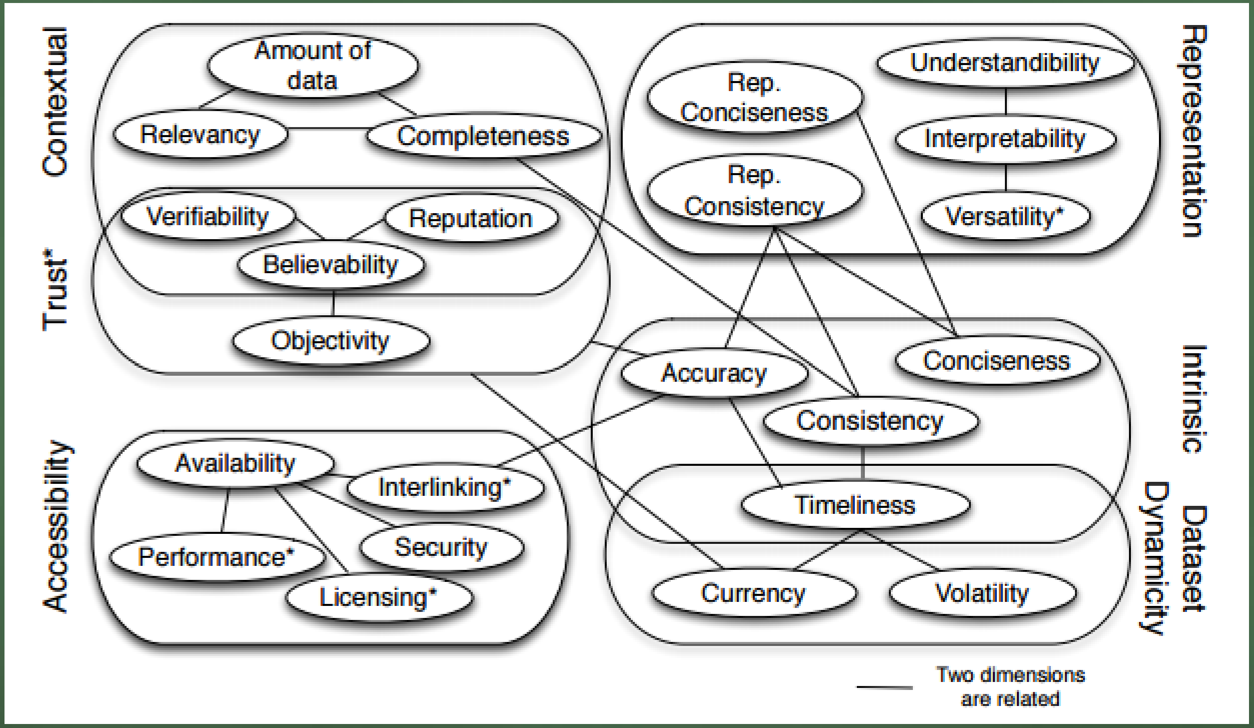
Data quality assessment metrics (1)

- Data Consistency
- Data Deduplication
- Data Accuracy
- Data Completeness
- Data Currency, aims to identify the current values of entities represented by tuples in a database, and to answer queries with the current values

	FN	LN	CC	AC	phn	street	city	zip	salary	status
t_1:	Mike	Clark	44	131	null	Mayfield	NYC	EH4 8LE	60k	single
t_2:	Rick	Stark	44	131	3456789	Crichton	NYC	EH4 8LE	96k	married
t_3:	Joe	Brady	01	908	7966899	Mtn Ave	NYC	NJ 07974	90k	married
t_4:	Mary	Smith	01	908	7966899	Mtn Ave	MH	NJ 07974	50k	single
t_5:	Mary	Luth	01	908	7966899	Mtn Ave	MH	NJ 07974	50k	married
t_6:	Mary	Luth	44	131	3456789	Mayfield	EDI	EH4 8LE	80k	married

Data quality assessment metrics (2)

- Zaveri[1] surveyed 96 metrics and categorized them into 6 dimensions:
    - Accessibility
    - Intrinsic
    - Trust
    - Dataset dynamicity
    - Contextual
    - Representational



Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: A survey. Semantic Web Journal 7 (2016) 63-93

Data quality assessment metrics (2)-- Contextual

Dimension	Metric	Description	Type
Completeness	degree to which classes and properties are not missing	detection of the degree to which the classes and properties of an ontology are represented [5,15,42]	S
	degree to which values for a property are not missing	detection of no. of missing values for a specific property [5, 15]	O
	degree to which real-world objects are not missing	detection of the degree to which all the real-world objects are represented [5,15,26,42]	O
	degree to which interlinks are not missing	detection of the degree to which instances in the dataset are interlinked [22]	O
Amount-of-data	appropriate volume of data for a particular task	no. of triples, instances per class, internal and external links in a dataset [14,12]	O
	coverage	scope and level of detail [14]	S
Relevancy	usage of meta-information attributes	counting the occurrence of relevant terms within these attributes or using vector space model and assigning higher weight to terms that appear within the meta-information attributes [5]	S
	retrieval of relevant documents	sorting documents according to their relevancy for a given query [5]	S

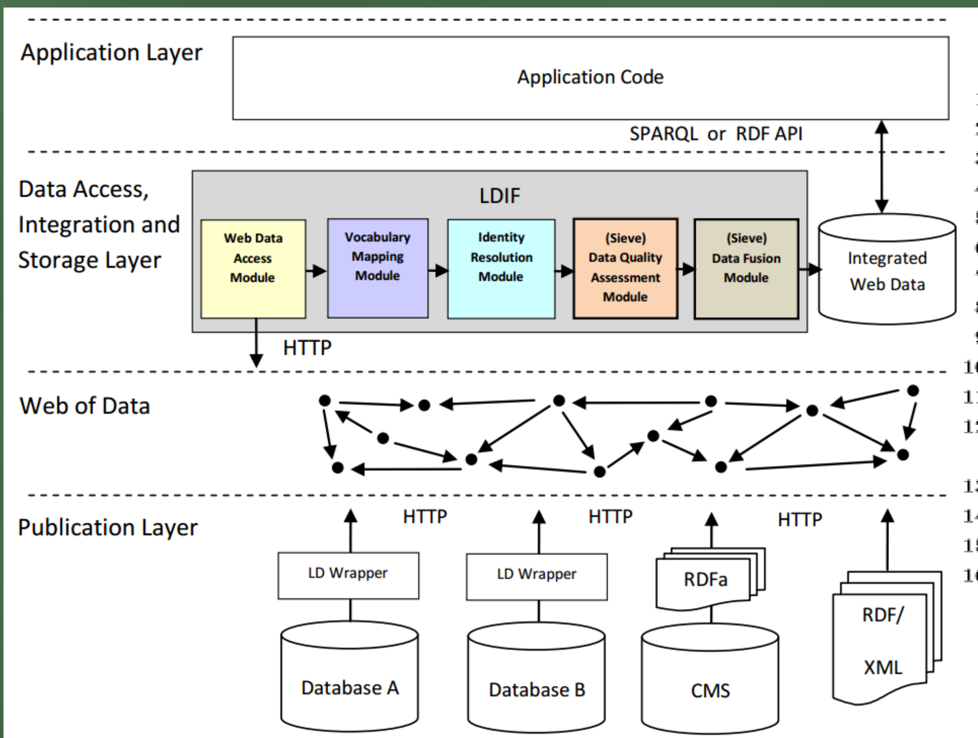
Data quality assessment tool (1)--Flemming

- Flemming provides a simple web user interface and a walk through guide that helps a user to assess data quality of a resource using a set of defined metrics
- Metric weight can be customizable either by user assignment, though no new metric can be added to the system.

Flemming, A. Quality characteristics of linked data publishing datasources. MA thesis. Humboldt-Universität zu Berlin, Institut für Informatik, 2011.

Data quality assessment tool (2)--Sieve

- In Sieve, metadata about named graphs is used in order to assess data quality, where assessment metrics are declaratively defined by users through an XML configuration



LDIF Architecture

```
<Sieve>
  <QualityAssessment>
    <AssessmentMetric id="sieve:recency">
    <ScoringFunction class="TimeCloseness">
        <Param name="timeSpan" value="7"/>
        <Input path="?GRAPH/provenance:lastUpdated"/>
    </ScoringFunction>
    </AssessmentMetric>
    <AssessmentMetric id="sieve:reputation">
      <ScoringFunction class="ScoredList">
        <Param name="priority"
            value="http://pt.wikipedia.org http://en.
                wikipedia.org"/>
      </ScoringFunction>
    </AssessmentMetric>
  </QualityAssessment>
</Sieve>
```

Sieve Data Quality Assessment Specification Example

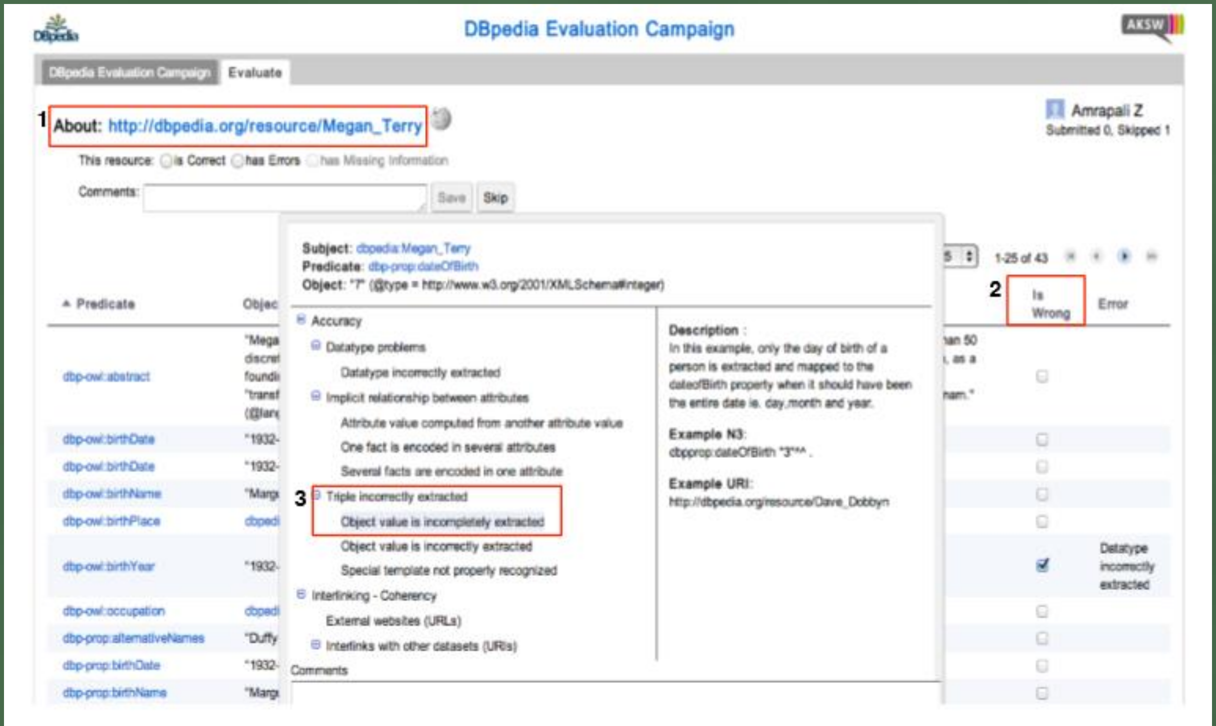
Mendes, P.N., M•uhleisen, H., Bizer, C.: Sieve: linked data quality assessment and fusion. In: Proceedings of the 2012 Joint EDBT/ICDT Workshops, ACM (2012) 116-123

Data quality assessment tool (3) -- TripleCheckMate

- TripleCheckMate is a crowdsourcing quality assessment tool focused on the correctness evaluation of DBpedia.
- The tool is released as open source under the Apache License. A successful use case of the tool can be seen in the Dbpedia Evaluation Campaign where 58 users evaluated a total of 792 resources (521 distinct) and 2928 distinct triples

Kontokostas, D., Zaveri, A., Auer, S., Lehmann, J.: Triplecheckmate: A tool for crowdsourcing the quality assessment of linked data. Knowledge Engineering and Semantic Web (2013) 265-272

Data quality assessment tool (3)-- TripleCheckMate



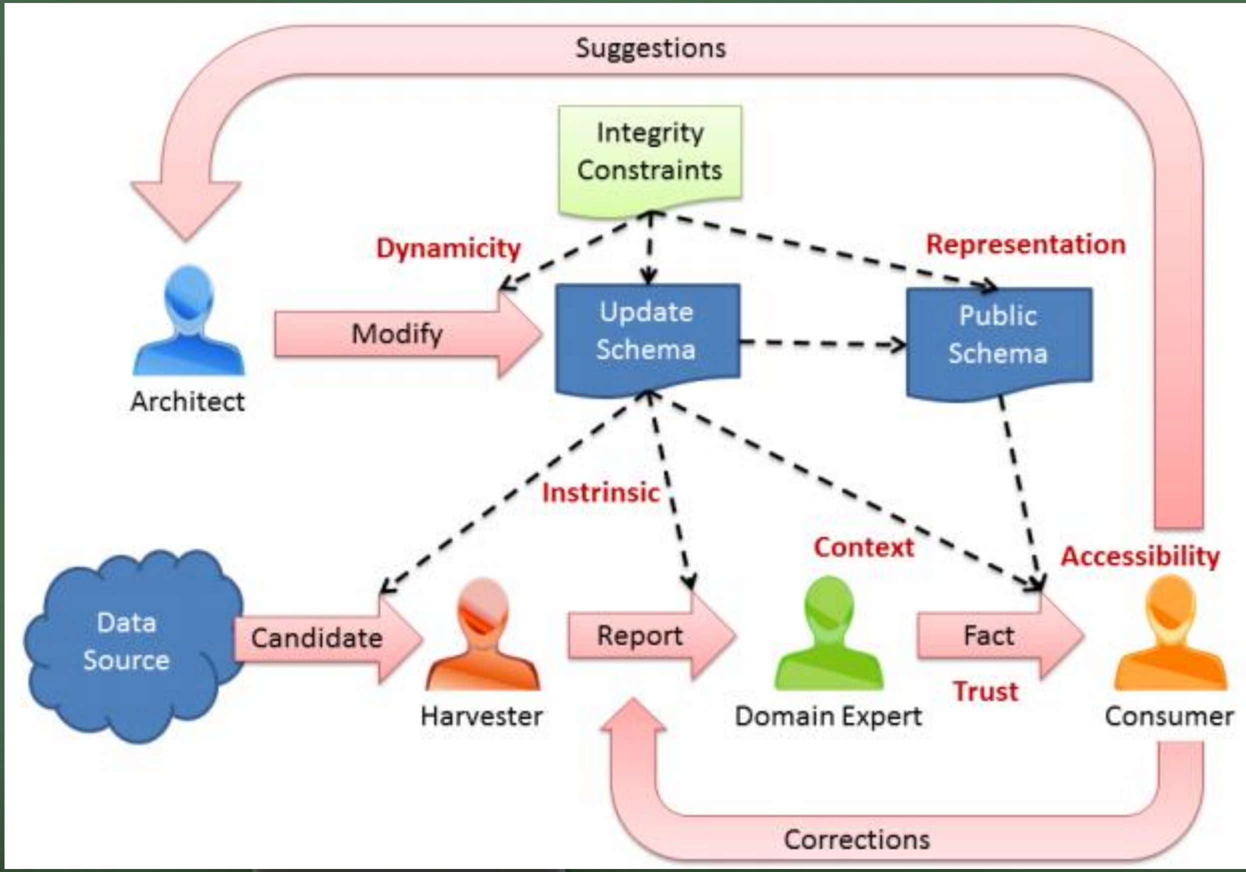
1. First, a user chooses a resource. 2. she is displayed with all triples belonging to that resources and evaluates each triple individually to detect quality problems. 3. If she finds a problem, she marks it and associates it with a relevant problem category.

Data quality assessment tool (3)--TripleCheckMate

Dimension	Category	Sub-category	DBpedia specific
Accuracy	Triple incorrectly extracted	Object value is incorrectly extracted	–
		Object value is incompletely extracted	–
		Special template not properly recognized	✔
	Datatype problems	Datatype incorrectly extracted	–
	Implicit relationship between attributes	One fact encoded in several attributes	✔
		Several facts encoded in one attribute	–
		Attribute value computed from another attribute value	✔
Relevancy	Irrelevant information extracted	Extraction of attributes containing layout information	✔
		Redundant attribute values	–
		Image related information	✔
		Other irrelevant information	–
Represensational-Consistency	Representation of number values	Inconsistency in representation of number values	–
Interlinking	External links	External websites	–
	Interlinks with other datasets	Links to Wikimedia	–
		Links to Freebase	–
		Links to Geospecies	–
		Links generated via Flickr wrapper	–

Data quality assessment tool (4)-- DaCura

- DaCura is a semi-automated process, framework and tools for harvesting, assessing, improving and maintaining high-quality linked-data

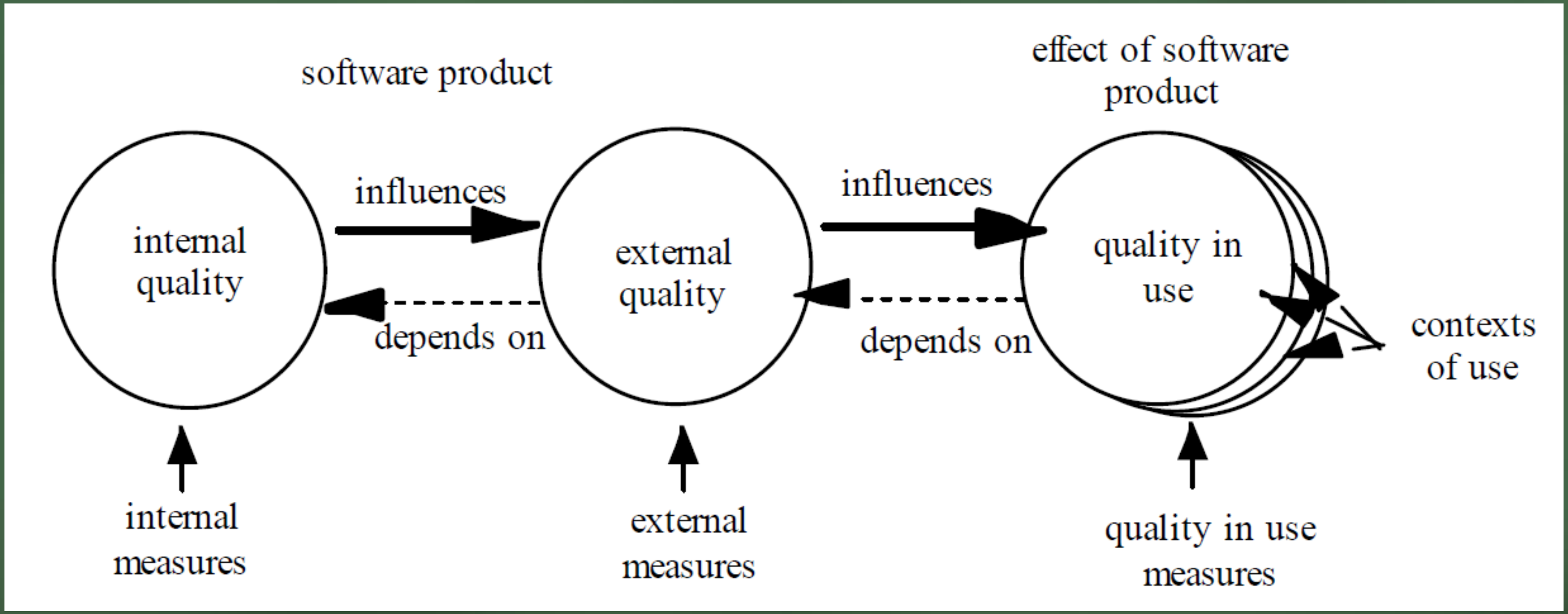


Schema-based Data quality assessment

Feeney K C, O'Sullivan D, Tai W, et al. Improving curated web-data quality with structured harvesting and assessment[J]. International Journal on Semantic Web and Information Systems (IJSWIS), 2014, 10(2): 35-62.

Data quality assessment standard -- ISO/IEC 25010

- **ISO/IEC 25010:2011** System and software quality models
- The quality in use model in ISO/IEC 25010[1] distinguishes three quality approaches: internal quality, external quality, and quality in use



Data quality assessment standard -- ISO/IEC 25012

- **ISO/IEC 25012:2008** Software engineering -- Software product Quality Requirements and Evaluation (SQuaRE) -- Data quality model
- ISO/IEC 25012:2008 categorizes quality attributes into fifteen characteristics considered by two points of view: **inherent** and **system dependent**.

Table 1 — Data quality model characteristics

Characteristics	*DATA QUALITY*	
	Inherent	*System dependent*
Accuracy	X	
Completeness	X	
Consistency	X	
Credibility	X	
Currentness	X	
Accessibility	X	X
Compliance	X	X
Confidentiality	X	X
Efficiency	X	X
Precision	X	X
Traceability	X	X
Understandability	X	X
Availability		X
Portability		X
Recoverability		X

Data quality assessment standard -- ISO/IEC 25024

- **ISO/IEC 25024:2015** Measurement of data quality
- ISO/IEC 25024:2015 defines data quality measures for quantitatively measuring the data quality in terms of characteristics defined in ISO/IEC 25012. It contains:
  - a basic set of data quality measures for each characteristic;
  - a basic set of target entities to which the quality measures are applied during the data-life-cycle;
  - an explanation of how to apply data quality measures;
  - a guidance for organizations defining their own measures for data quality requirements and evaluation.

ID	Name	Description	Measurement function	DLC Target entities Properties
Acc-I-1	Syntactic data accuracy	Ratio of closeness of the data values to a set of values defined in a domain	X=A/B A= number of data items which have related values syntactically accurate B= number of data items for which syntactic accuracy is required	All DLC except Data design Data file Data item, data value
NOTE 1 A single value is considered "syntactically accurate" when it is the same as one from an identified source of validated information: the result is "yes" or "no". NOTE 2 An example of a low degree of syntactic accuracy is when the word Mary is stored as Marj. NOTE 3 See ISO/IEC 25012, 5.3.1.1.				
Acc-I-2	Semantic data accuracy	Ratio of how accurate are the data values in terms of semantics in a specific context	X=A/B A= number of data values semantically accurate B= number of data values for which semantic accuracy is required	All DLC except Data design Data file Data value
NOTE 1 A single value is considered "semantically accurate" when the meaning (the content) corresponds to the reality. NOTE 2 An example of a low degree of semantic accuracy is when the name of John is recorded instead of George; both names are syntactically accurate, so only George is semantically accurate. NOTE 3 See ISO/IEC 25012, 5.3.1.1.				
Acc-I-3	Data accuracy assurance	Ratio of measurement coverage for accurate data	X=A/B A= number of data items measured for accuracy B= number of data items for which measurement is required for accuracy	All DLC except Data design Data file Data item

Challenges in data quality assessment

How to evaluate the quality when

- Data are different
- Usage for the data are different.
- Data volume is tremendous.
- Data is changing
- Large amount of metrics to choose from

L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era",Data Sci. J., vol. 14, no. 2, pp. 1-10, 2015.

PART II
TESTING AND ASSESSING THE DATA OF KNOWLEDGE GRAPHS

- An Introduction of Data Quality Problem
- Models and Metrics of Data Quality Assessment
- Tools and Standards of Data Quality Assessment
- Challenges in Data Quality Assessment
- Some Recent Results

Some recent results

- Data quality assessment standard in big data Trading (on-going)
- Data quality assessment on internal quality
- Data quality assessment on quality in use
- Data quality assessment on EMR (on-going)

Data quality assessment standard in big data trading

ICS 点击此处添加 ICS 号

点击此处添加中国标准文献分类号

中华人民共和国国家标准

GB/T —XXXX

信息技术 数据交易 数据质量评估

Information technology-data transaction-

Data Quality Assessment

点击此处添加与国际标准一致性程度的标识

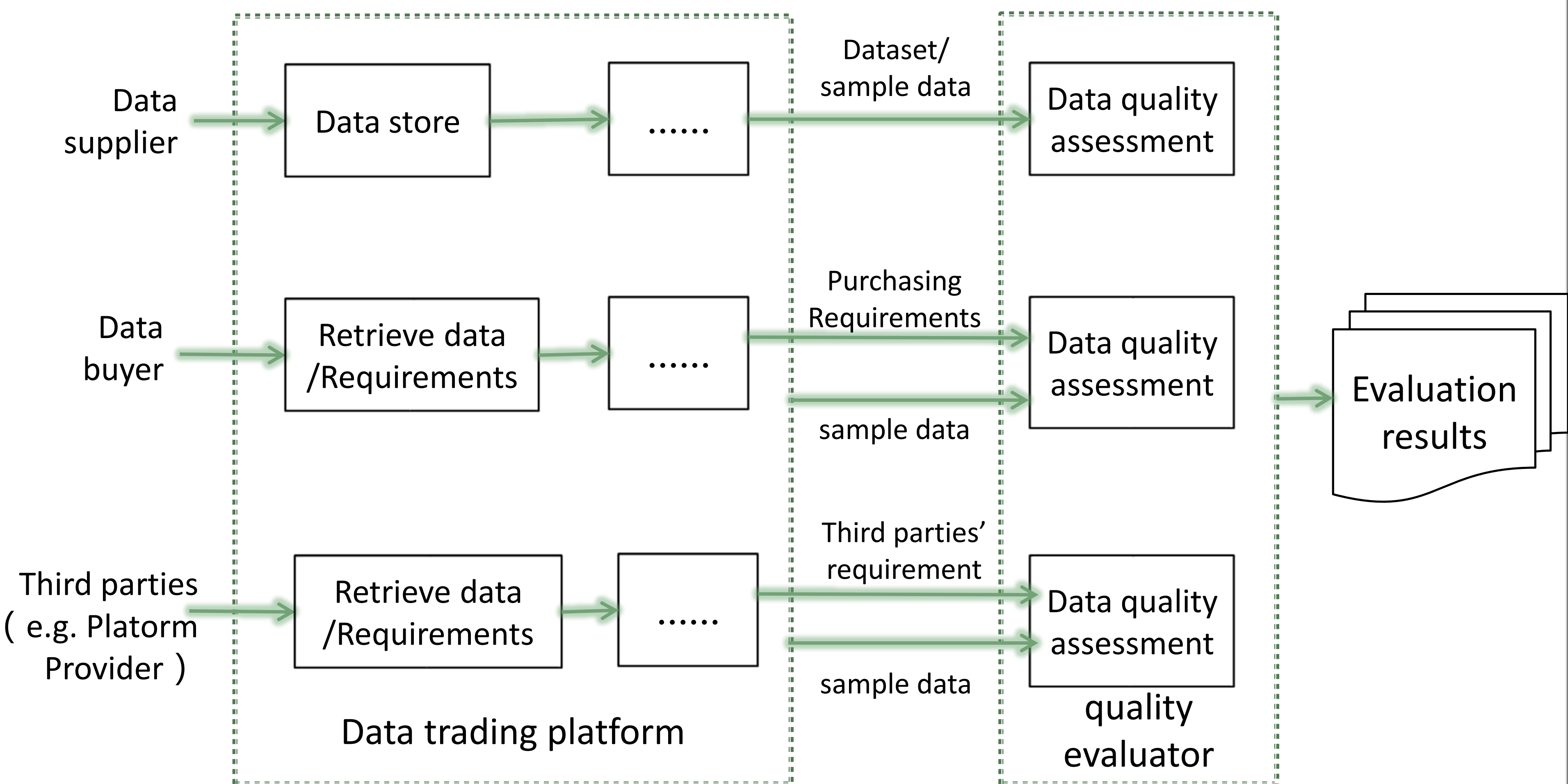
（工作组讨论稿）

XXXX-XX-XX发布　　　　XXXX-XX-XX实施

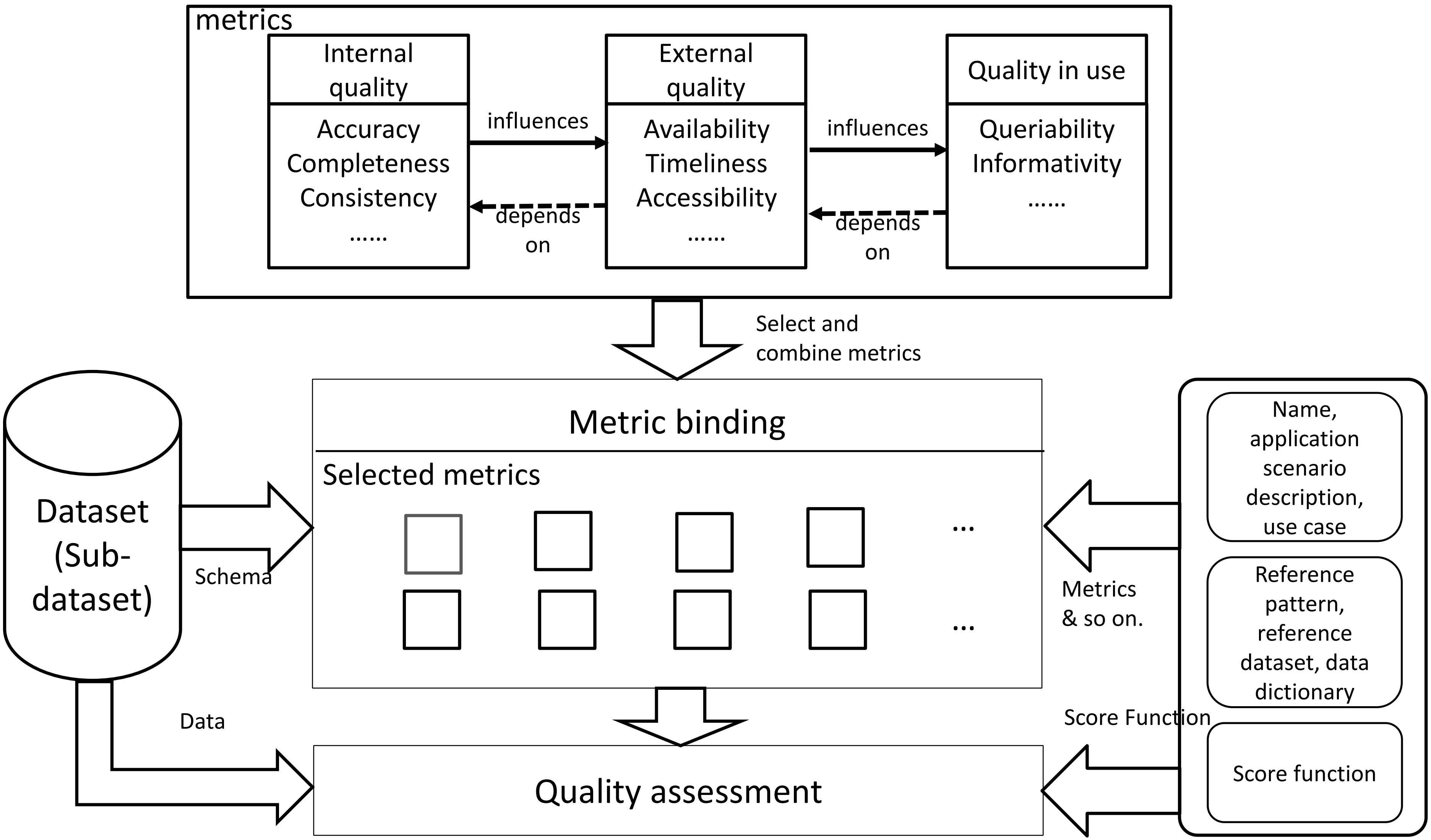
中华人民共和国国家质量监督检验检疫总局
中国国家标准化管理委员会　发布

目　次

Application scenario in data quality assessment standard in big data trading

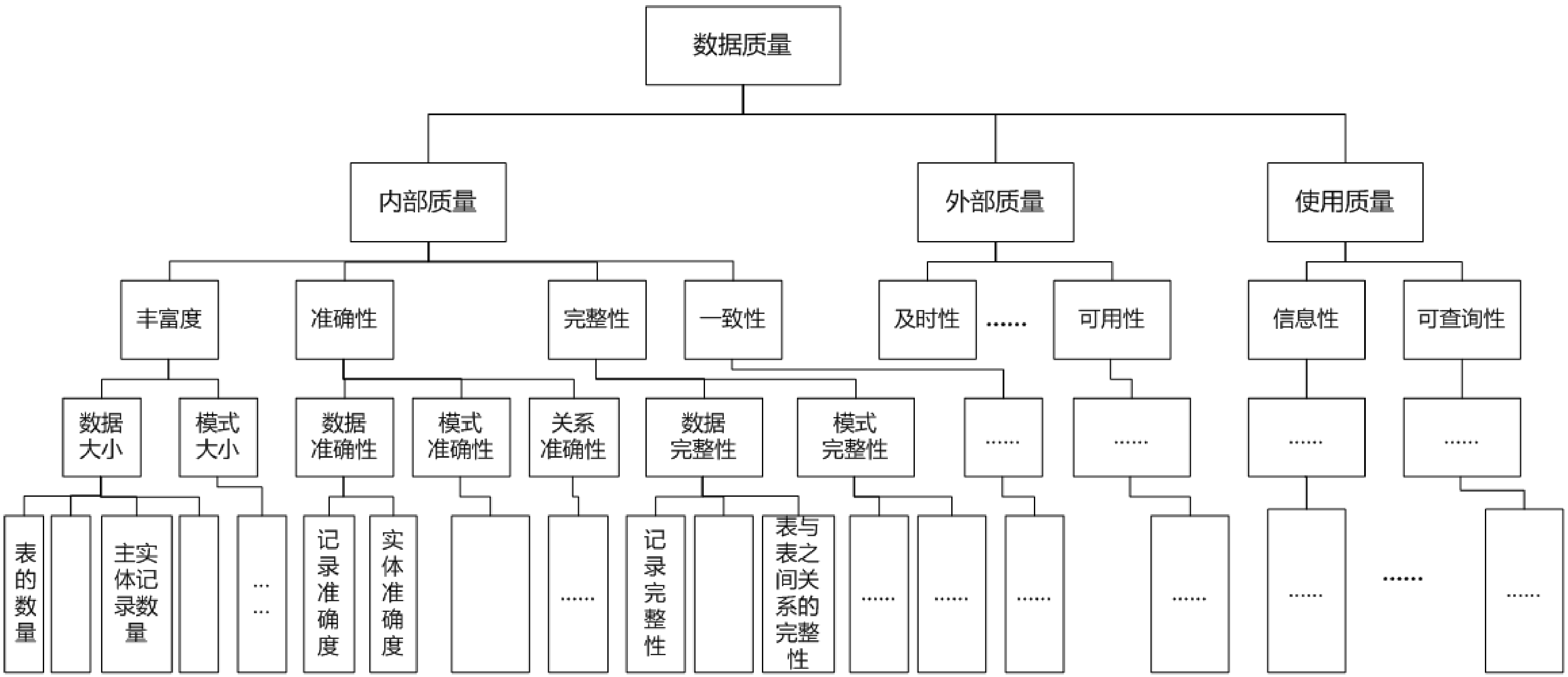


Main framework of assessment standard



Quality metric framework of assessment standard

- Metric framework contains four layers：



Definition of the quality characteristics

- Defined quality characteristics ：
    - Accuracy, Completeness, Consistency, etc
    - referenced from ISO/IEC 25012
- New quality characteristics ：
    - Richness, Informativity, Queriability, etc

Definition and computation of quality metrics

(1) metric computation -- no binding to usage context

- e.g. the metrics in richness: the number of tables/concepts
  - Definition: the number of tables/concepts in a dataset
  - Computation: directly acquire by SQL or SPARQL

(2) metric computation – metric binding to usage context

- eg. The metrics in completeness: the View Completeness
  - Definition:

    The completeness of the attributes in target tables/concepts, evaluated in Design view whether it missing required attributes or not

  - Computation: $ComR(X) = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{NOIP_i}{NOR_i}\right)$

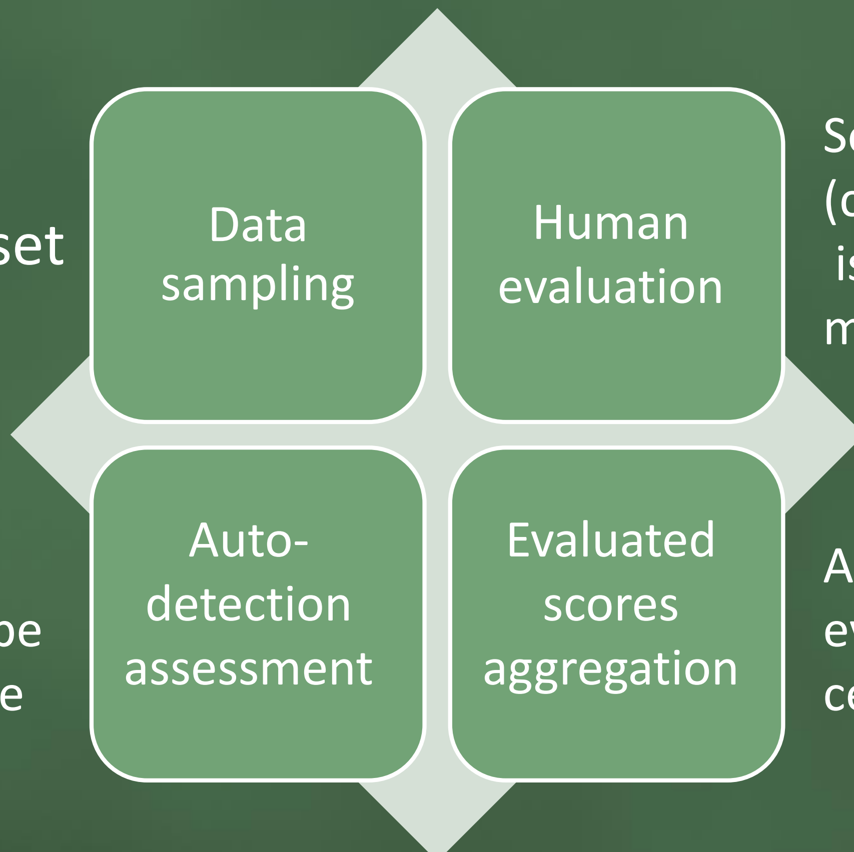
    NOR_i is the number of attributes in the Table/Concept i,

    $NOIP_i$ is the number of required but none-existent attributes in the Table/Concept i

Assessment method in assessment standard

Huge dataset

Some metrics (correctness) is difficult to evaluate by machine



Some special data errors (outlier) can be detected by machine

Aggregate the results of evaluated metrics to a centesimal score

Assessment process in assessment standard

Confirm evaluation content
确定评估内容

Obtain domain context
获取领域上下文

Select metrics and scoring standard
确定评估度量和评分标准

Sample data
数据抽样

Evaluate data
数据评估

Export evaluation report
报告编制

	评估内容
目的	a）确定具备预定目标的一个或多个数据集 b）确定具备预定目标的一个或多个主实体 c）确定具备预定目标的一个或多个主属性
用途	a) 确定具备使用范围的一个或多个数据集 b) 确定具备使用范围的一个或多个主实体 c) 确定具备使用范围的一个或多个主属性
用户自定义	a）数据供方自定义的待评估数据集、主实体、主属性 b）数据需方定制要求的待评估数据集、主实体、主属性

下列方式获取： 符合下

料引用

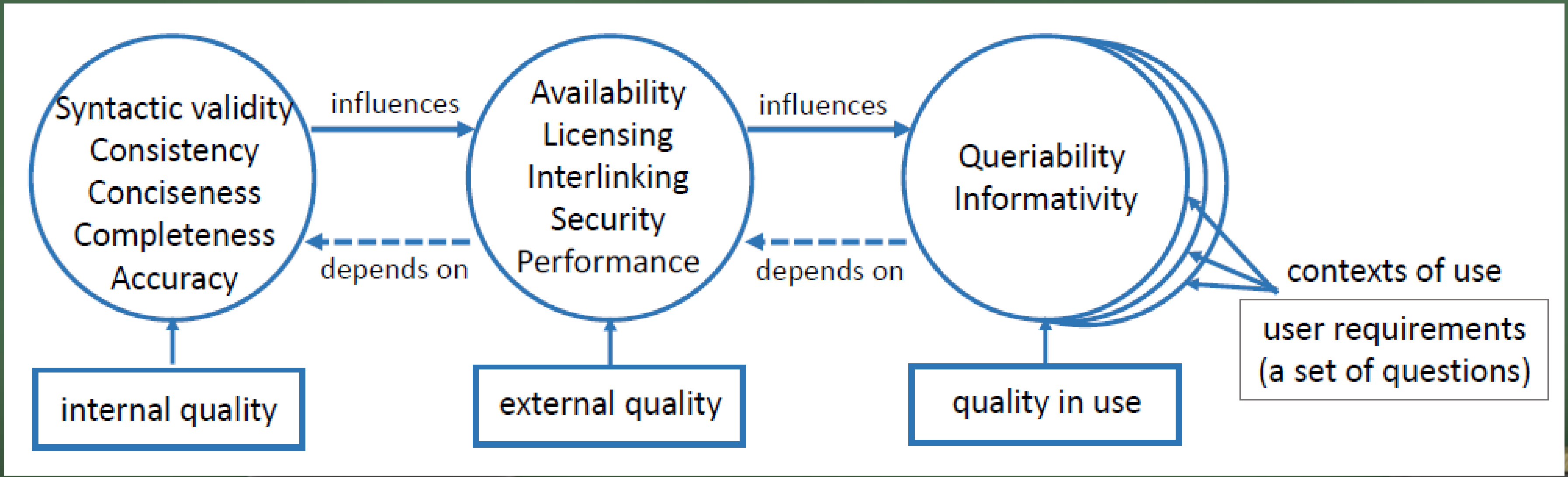
分

Some recent results

- Data quality assessment standard in big data transaction
- Data quality assessment on internal quality
- Data quality assessment on quality in use
- Data quality assessment on EMR (on-going)

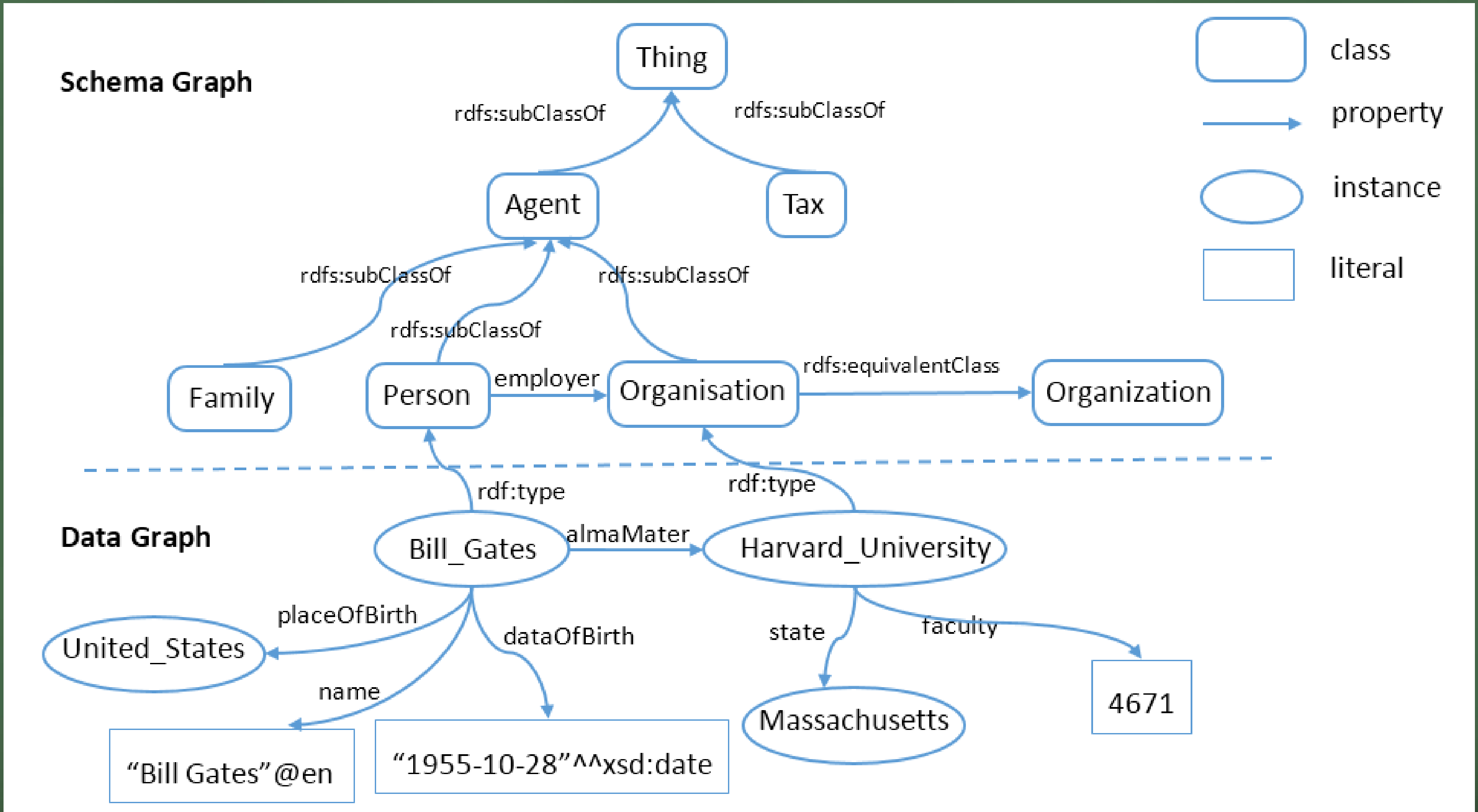
Data quality model

- Inspired by the software quality model in ISO/IEC 25010, we also divide the quality of LOD into three factors, namely, internal quality, external quality, and quality in use
- We further map the quality dimensions of linked data mentioned in Zaveri[1] into internal and external factors in our model



 Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: A survey. Semantic Web Journal 7 (2016) 63-93

Conceptual Representation



Metrics Definition — Richness Dimension

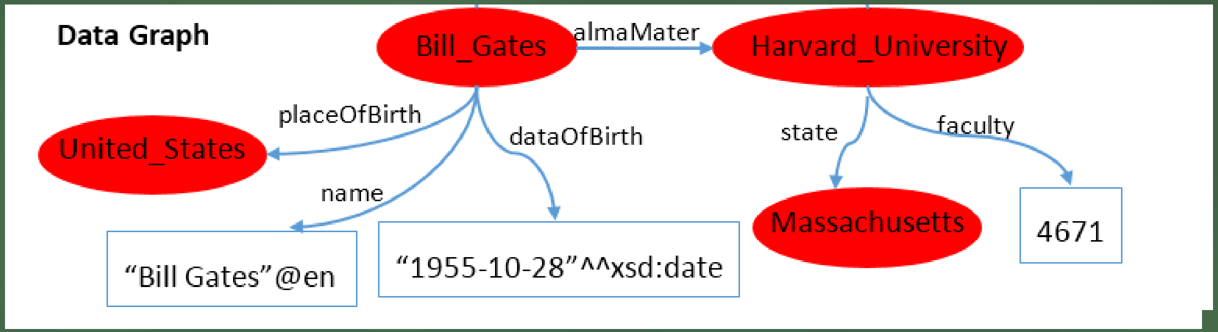
Metric Set	Category	Metrics	Description	
Richness	Data Size	Number of Instances (NOI)	Total number of instances in KB	$\|N_i\|$
		Number of Facts (NOF)	Total number of facts in KB	$\|E_s\|$
	Schema Size	Number of Classes (NOCL)	Total number of classes in KB	$\|N_s\|$
		Number of Categories (NOC)	Total number of categories in KB $noc = \{C \mid (i, dct{:}subject,\ C) \in E_d\}$	$\|noc\|$
		Number of Properties (NOP)	Total number of properties in KB	$\|P_d\|$
	Degree of Network	Average Fact Numbers (AFN)	Average fact number per Instance	$\|E_d\| / \|N_i\|$
		Zero-out Instances (ZOI)	Percentage of instances having no facts $zoi = \{ i \mid out-degree(i) = 0,\ i \in N_i\}$	$\|zoi\| / \|N_i\|$
		Zero-in Instances (ZII)	Percentage of instances that are not property values of any instances. $zii = \{ i \mid in-degree(i) = 0,\ i \in N_i\}$	$\|zii\| / \|N_i\|$
		Instantiation Ratio	Percentage of instances which have rdf:type relations to classes $ir = \{ i \mid (i,\ rdf{:}type,\ C),\ C \in N_s,\ i \in N_i\}$	$\|ir\|/\|N_i\|$
	Data Size over Domain	Number of Instances over Domain (NOID)	Total number of instances in particular domain C, $C \in N_s$ $i(C) = \{ i \mid (i, rdf{:}type, c), c \in C_s, C_s$ is the set of descendant classes of C $\}$	$\|i(C)\|$
		Number of Facts over Domain (NOPD)	Total number of facts in particular domain C $f(C) = \{ (i,\ p,\ n) \mid (i,\ p,\ n) \in E_d,\ i \in i(C)\}$	$\|f(C)\|$

Metrics Defined on Overlapped Instances

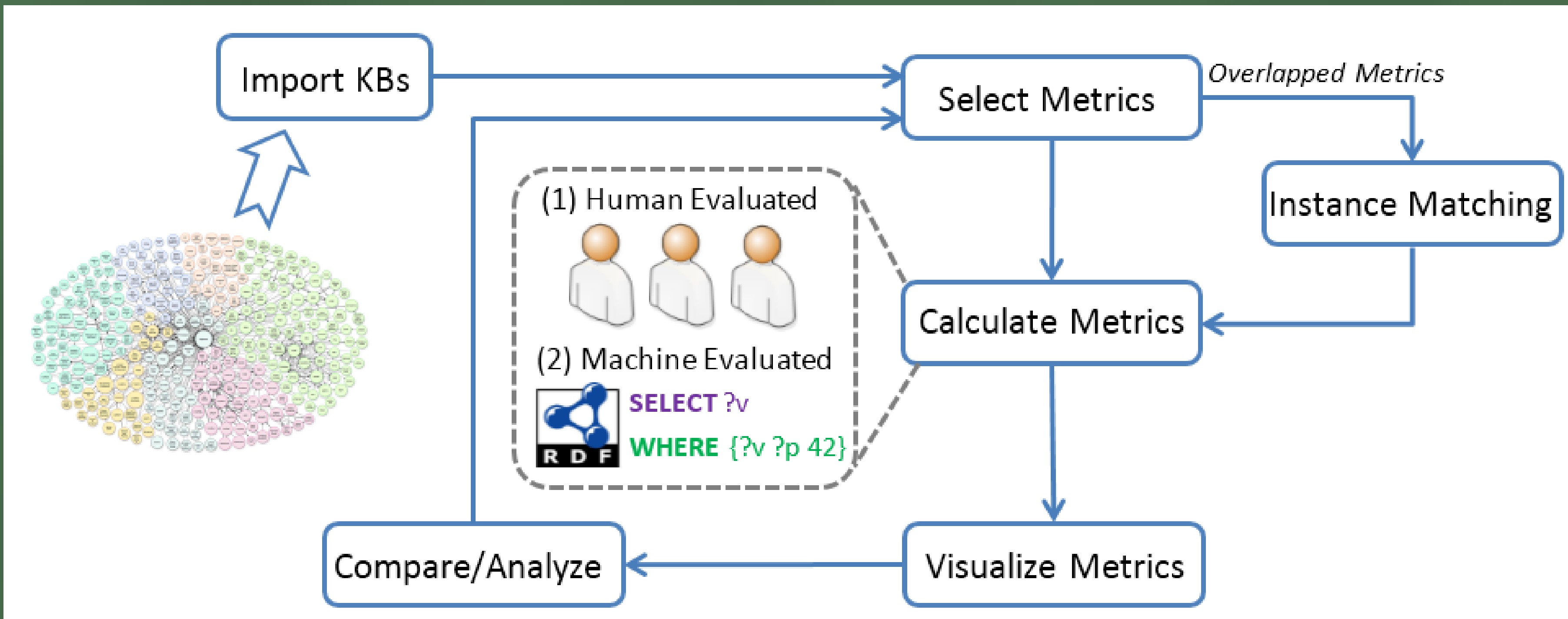
Metric Set	Category	Metrics	Description
Comparative metrics	Data Richness Comparison	Overlapped Instances (OI)	Number of overlapped instances for some KBs $\lvert \mathrm{I}_1 \rvert$
		Overlapped Instances over Domain (OIOD)	Number of overlapped instances in KB K for particular domain C $\mathrm{oiod(K,C)} = \{\mathrm{i} \mid \mathrm{i} \in \mathrm{I}_k, (\mathrm{i},\ \mathrm{rdf{:}type},\ \mathrm{C}) \in E_d\}$ $\lvert \mathrm{oiod(K,C)} \rvert$
		Facts of Overlapped Instances (FOI)	Number of facts of overlapped instances for KB K $\mathrm{foi(K,C)} = \{(\mathrm{i},\ \mathrm{p},\ \mathrm{n}) \mid (\mathrm{i},\ \mathrm{p},\ \mathrm{n}) \in E_d,\ \mathrm{i} \in \mathrm{I}_k\}$ $\lvert \mathrm{foi(K,C)} \rvert$
	Data Correctness Comparison	Correctness Ratio of Facts over Overlapped Instances (CRFO)	The percentage of correct facts of overlapped Instances. $1 - \frac{The\ number\ of\ wrong\ facts}{The\ number\ of\ all\ sampled\ facts}$
		Correctness Divergence (CD)	Distribution divergence on correctness of facts over overlapped instance of different KBs.

Metric: Number of Instances

- Number of Instances (NOI)： Total number of instances in KB.
- NOI = |Ni|
- NOI = |Ni| = 4



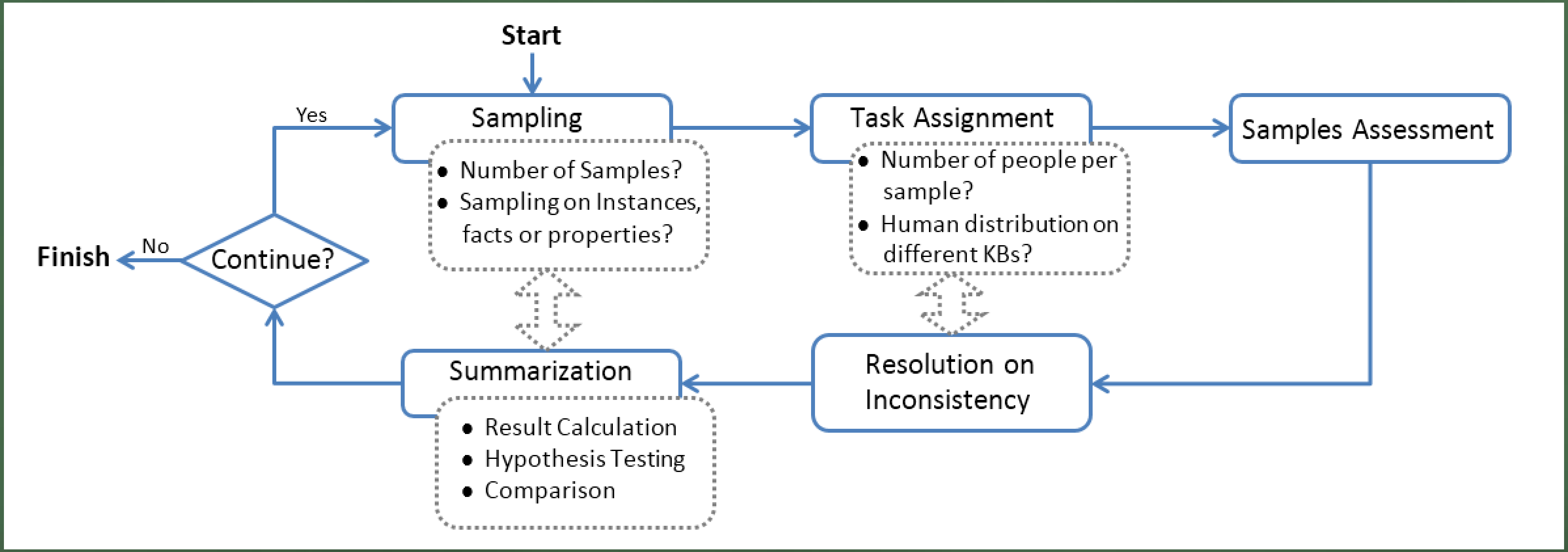
Overall Quality Evaluation Process



Tong Ruan, Xu Dong, Haofen Wang, and Yang Li. Evaluating and Comparing Web-Scale Extracted Knowledge Bases in Chinese and English. JIST2015

Human Evaluation Process

- Metrics (Correctness) evaluated by human may require too much human efforts and suffer from individual subjectivity from evaluators

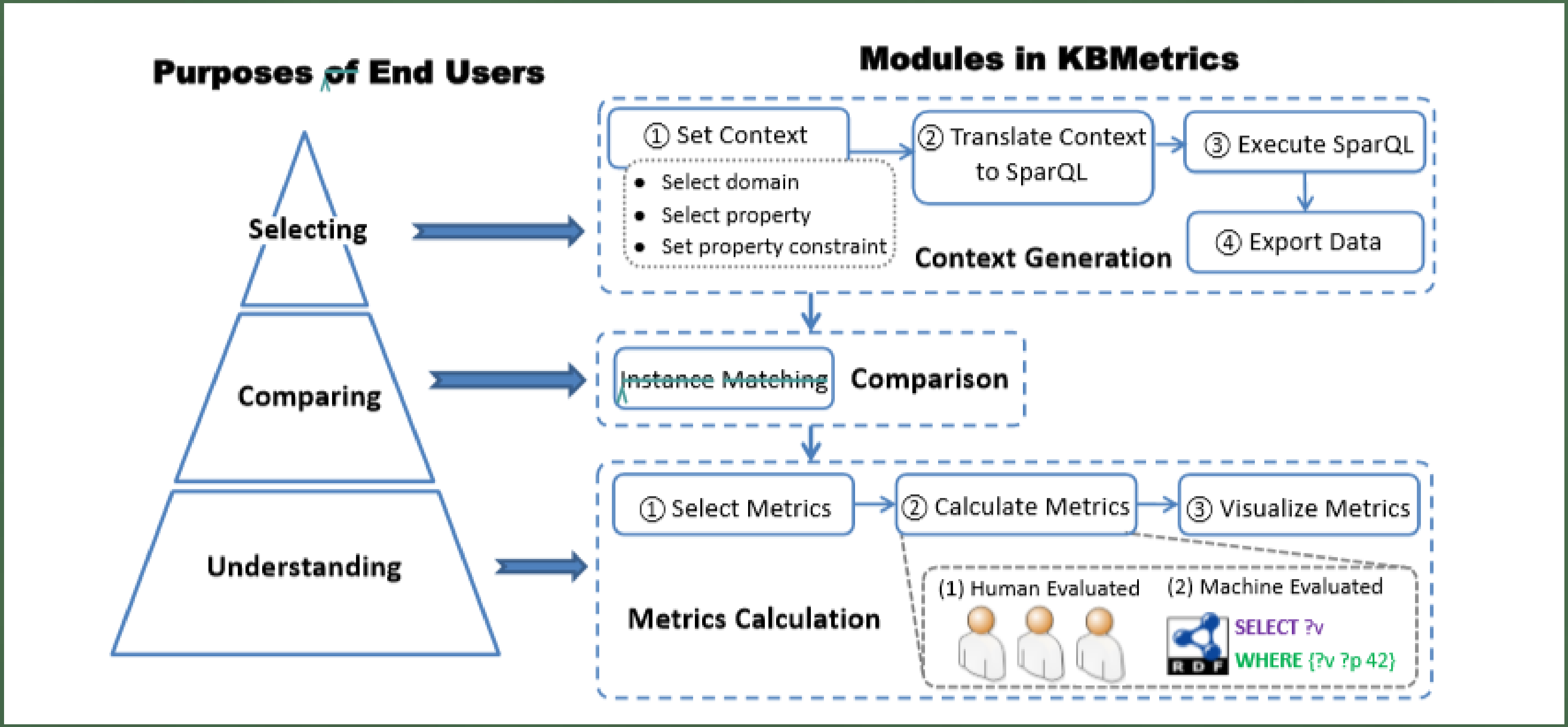


Metric: Sampling method--Random Sampling

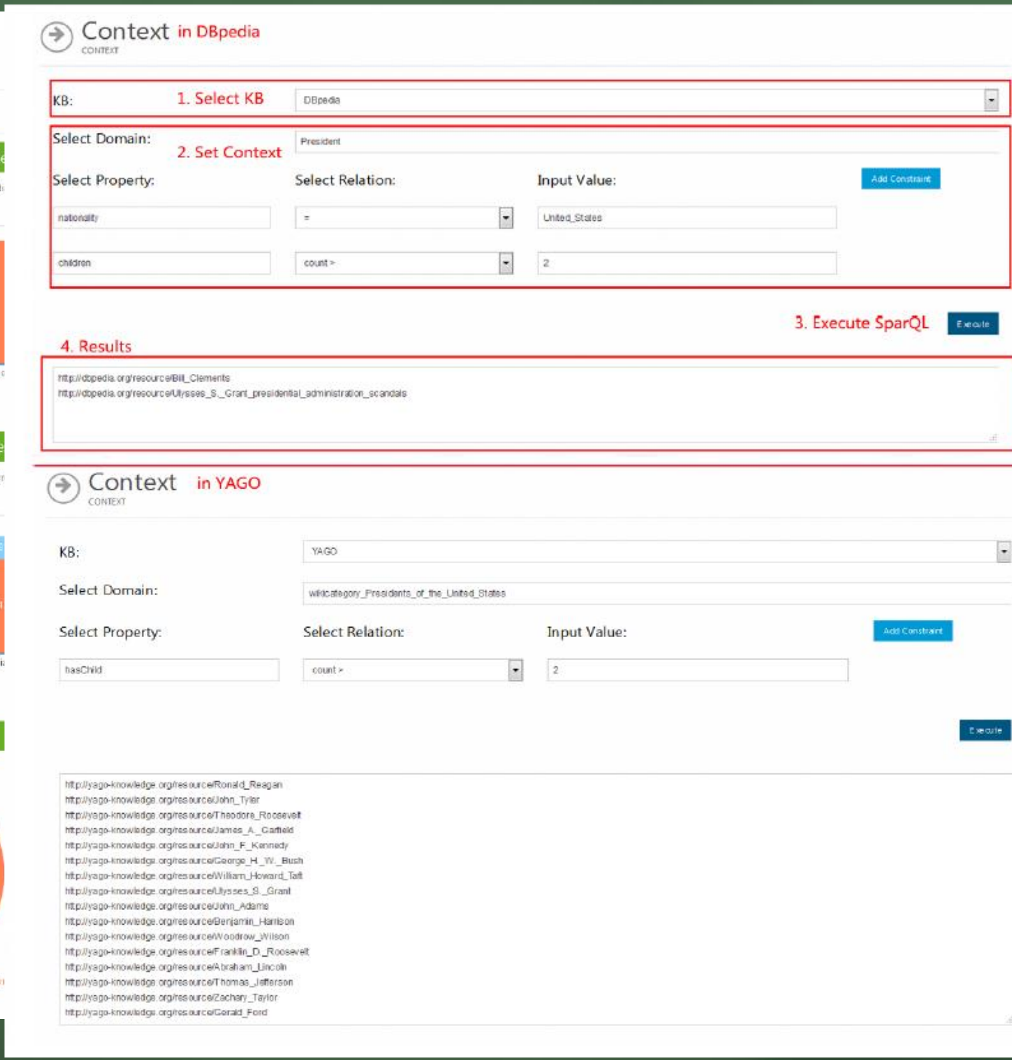
$$n = [(z^2 * p * q) + ME^2] \, / \, [ME^2 + z^2 * p * q/N]$$

- N is size of the data set, ME is the marginal of error
- It is assumed we do not have prior knowledge of the correctness ratio, and p is set to 0.5 in default
- User could input Alphas (1-condence level) for an estimation problem, and z can be derived from Alpha

Tools — KBMetrics (1)



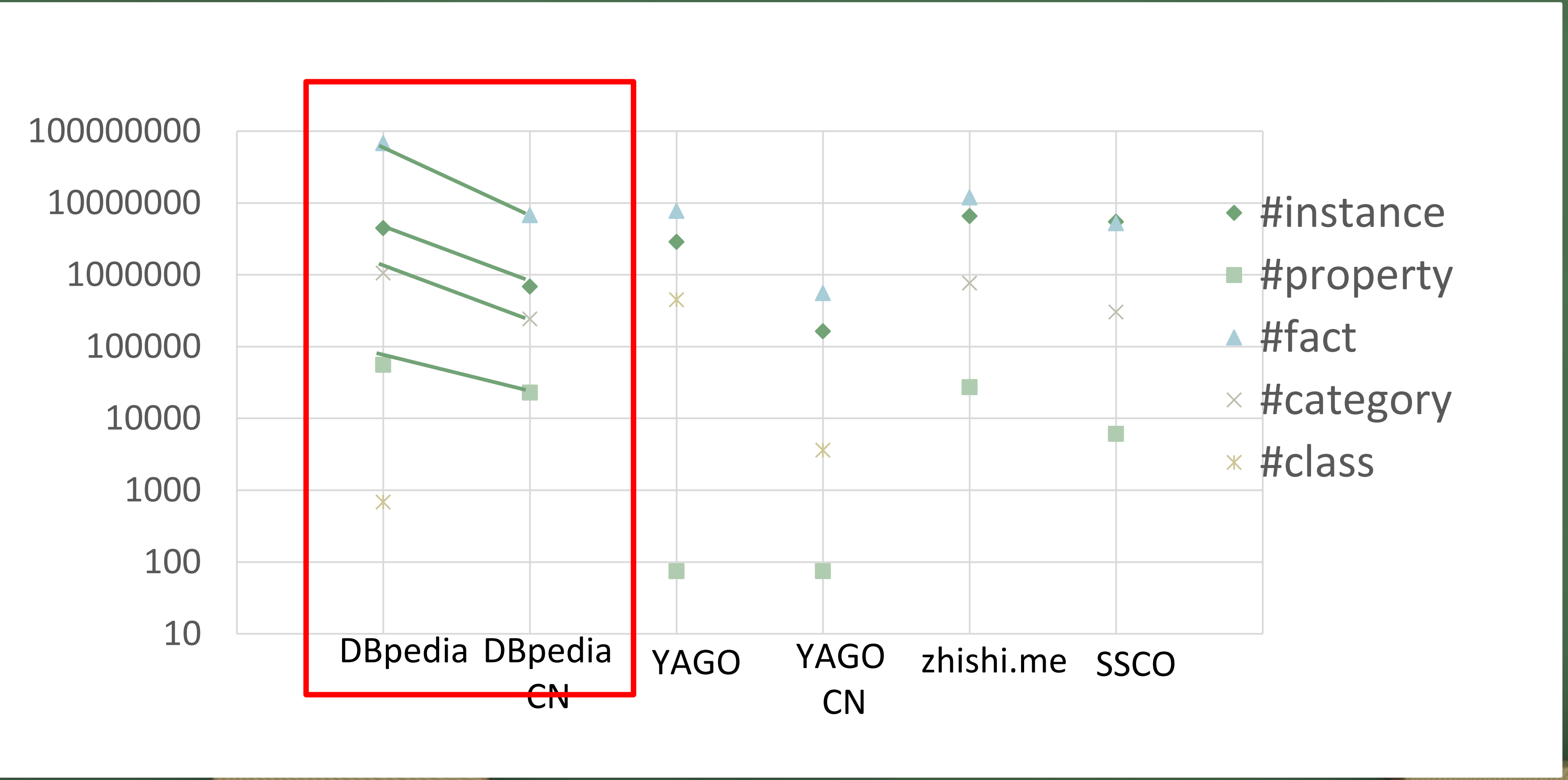
Tools — KBMetrics (2)



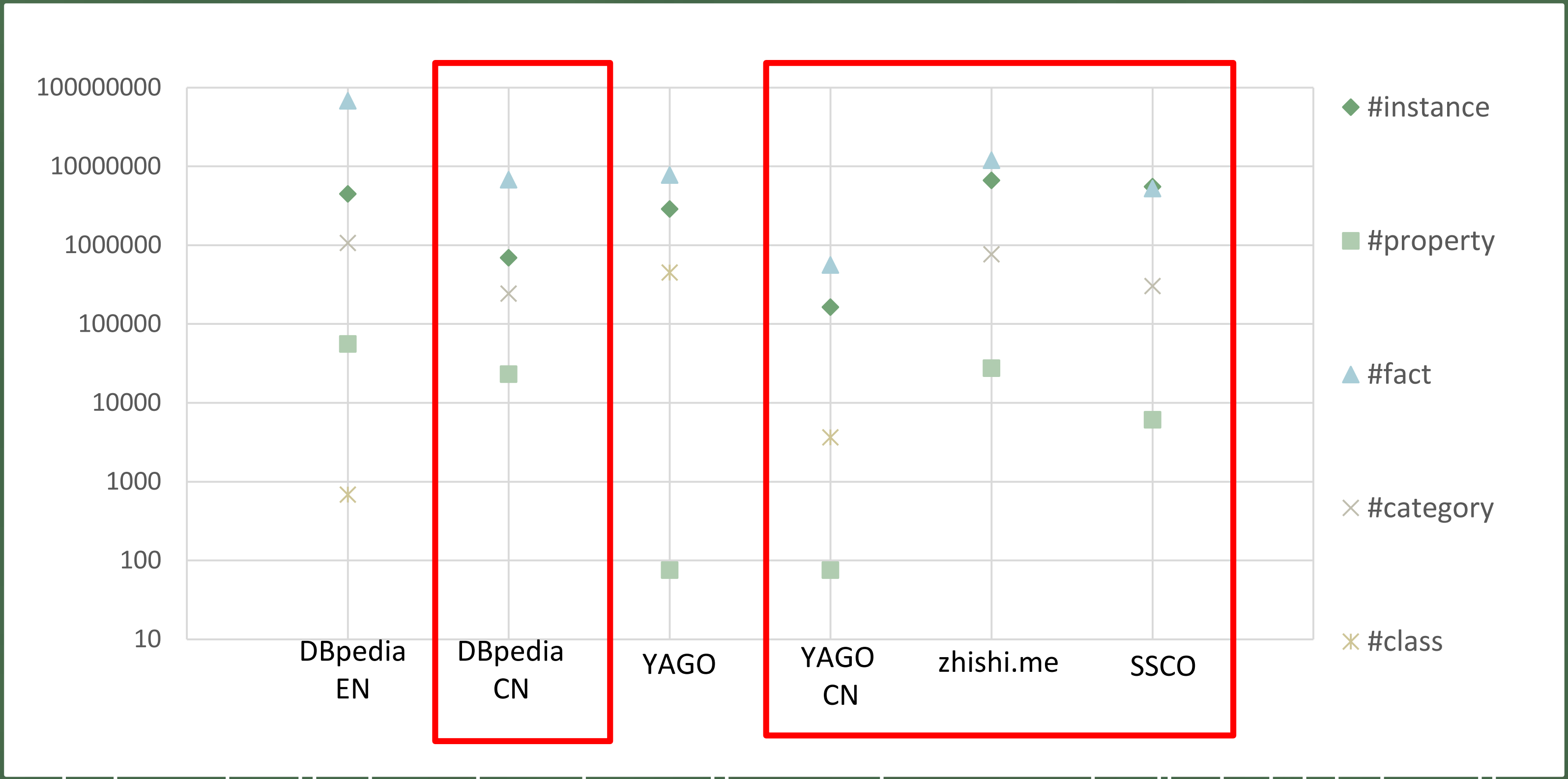
Evaluated Knowledge Bases

KB	Availability	License	Multi-language	InterLinks
DBpedia-EN(2014) DBpedia-CN(2014)	SparQL endpoint, dump	Creative Commons Attribution-ShareAlike License and GNU Free Documentation License	Localized versions of DBpedia are available in 125 languages	owl:sameAs
YAGO2s	SparQL endpoint, dump	Creative Commons Attribution 3.0 License	rdfs:label in localized languages	--
Zhishi.me	SparQL endpoint	?	Chinese	owl:sameAs
SSCO	--	?	Chinese	--

DBpedia-EN is richer than DBpedia-CN



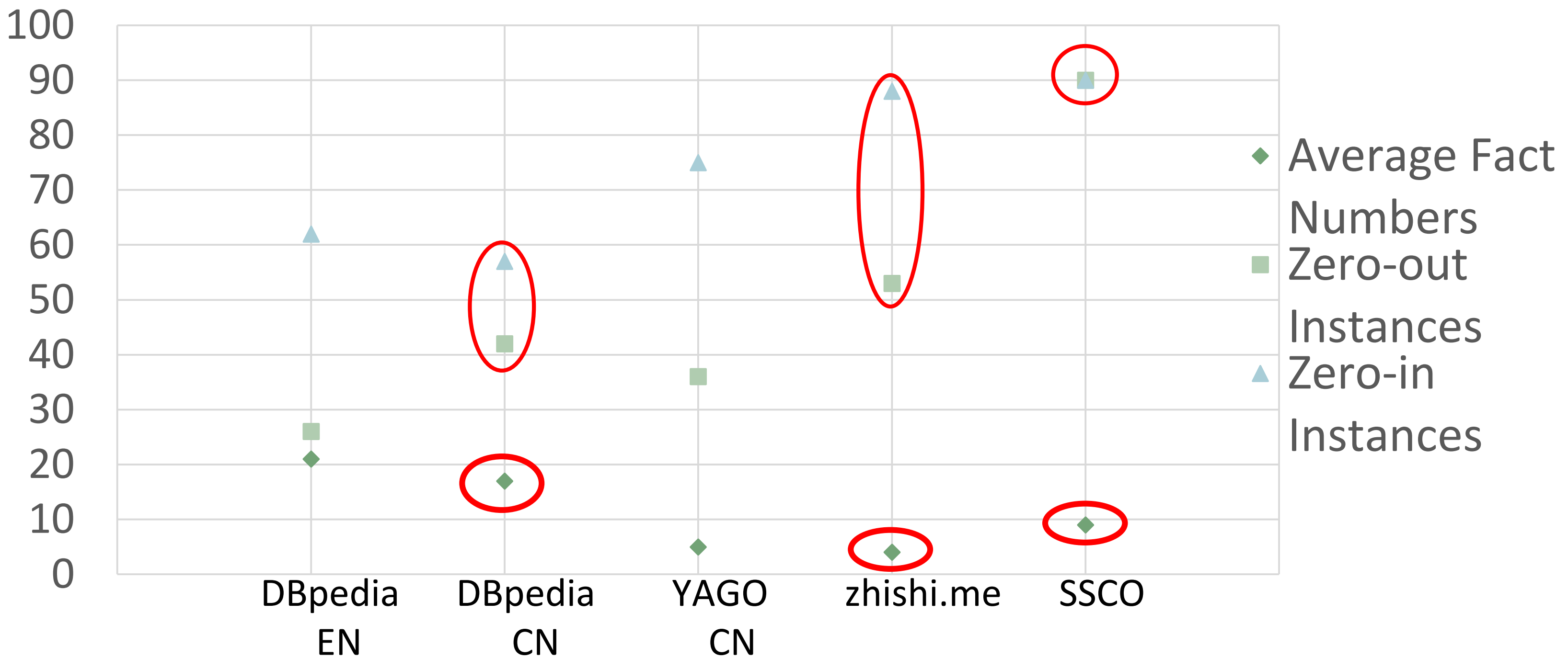
Zhishi.me has the largest number on data size and schema size



- zhishi.me has the largest number on those metrics, since it collected data from three Encyclopedic Web sites
- While the number of instance of SSCO is almost as much as that of zhishi.me, the number of fact of SSCO is less than DBpedia-CN, since SSCO filtered out low frequency properties
- The size of YAGO-CN in all aspects is very small, because only 5% of YAGO's instances contain Chinese label

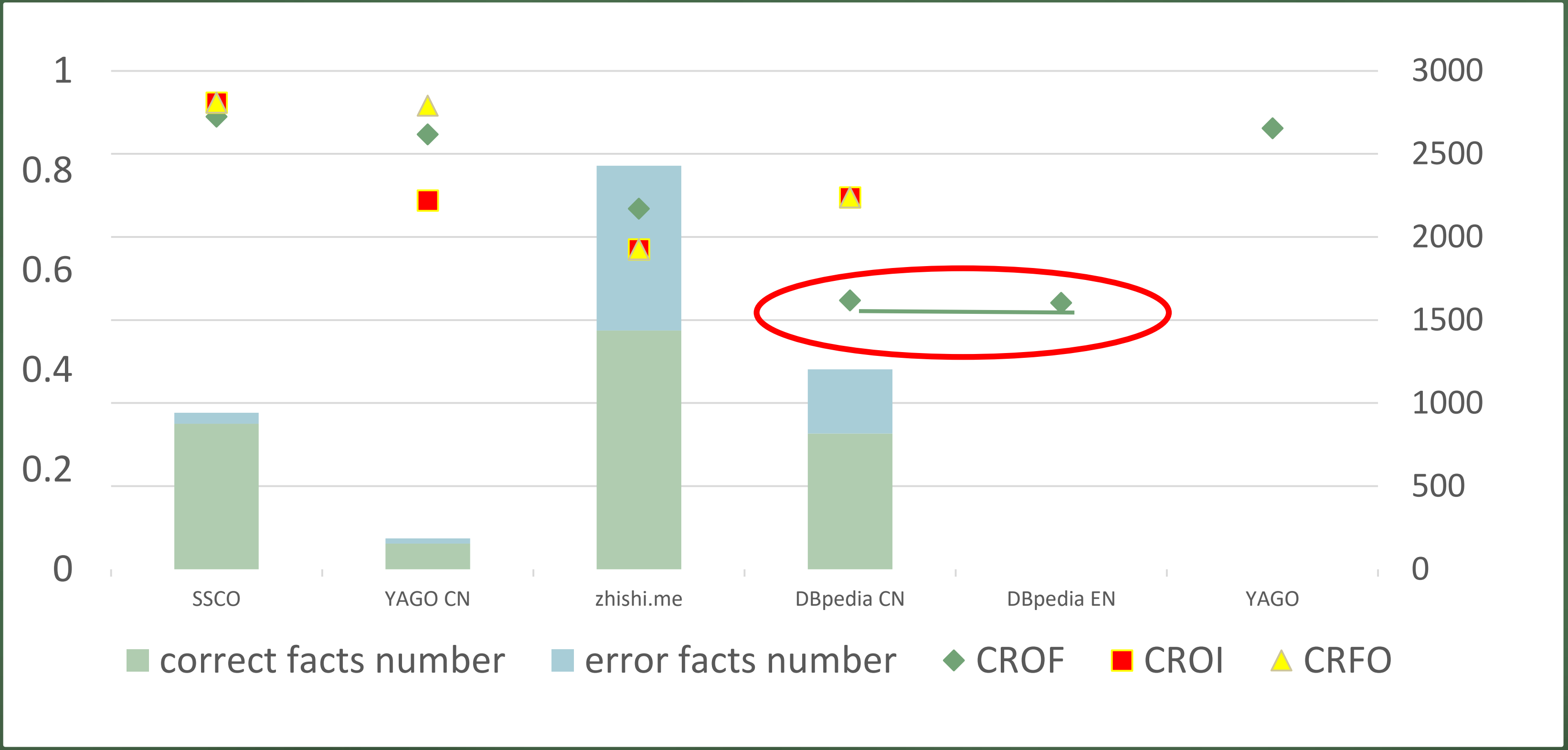
Instances in DBpedia-CN are linked to each other tighter than those in SSCO and zhishi.me

53% of the instances of zhishi.me have no facts and only 10% instances of SSCO have facts. In general, the instances in DBpedia-CN are linked to each other tighter than those in SSCO and zhishi.me.



The average fact number metric of zhishi.me and SSCO are smaller than that of DBpedia-CN. The reason may be that instances from Baidu-Baike and Hudong-Baike contain little facts.

DBpedia-CN and DBpedia-EN have the similar level of correctness statistically via hypothesis testing



Zhishi.me has the largest number of correct facts. SSCO has the highest correct ratio

Zhishi.me has the largest number of correct facts, while SSCO has the highest correct ratio. Since SSCO has very small number of facts, we infer that SSCO compares data from different sources and filters out the low frequency properties.

Error Analysis

—Error type: Lack of Context

- YAGO’s metafact is incomplete

#@ < id_vex9wm_1ul_xr0xdt>
<Paul_Lambert> <playsFor> <Scottish_League_XI>

< id_vex9wm_1ul_xr0xdt> <occursSince> "1990-##-##"^^xsd:date

Paul Lambert at Tallaght Stadium, August 2013

Personal information			
Full name	Paul Lambert[1]		
Date of birth	7 August 1969 (age 46)		
Place of birth	Glasgow, Scotland		
Height	1.81 m (5 ft 11 in)[2]		
Playing position	Midfielder		
Senior career*			
Years	**Team**	**Apps†**	**(Gls)†**
1986–1993	St. Mirren	227	(14)
1993–1996	Motherwell	103	(6)
1996–1997	Borussia Dortmund	44	(1)
1997–2005	Celtic	193	(14)
2005–2006	Livingston	7	(0)
Total		**574**	**(35)**
National team			
1990	Scottish League XI	1	(0)
1991–1992	Scotland U21[3]	5	(2)
1993–1996	Scotland B[4]	2	(0)
1995–2003	Scotland	40	(1)
Teams managed			
2005–2006	Livingston		
2006–2008	Wycombe Wanderers		
2008–2009	Colchester United		
2009–2012	Norwich City		
2012–2015	Aston Villa		

* Senior club appearances and goals counted for the domestic league only.

† Appearances (goals)

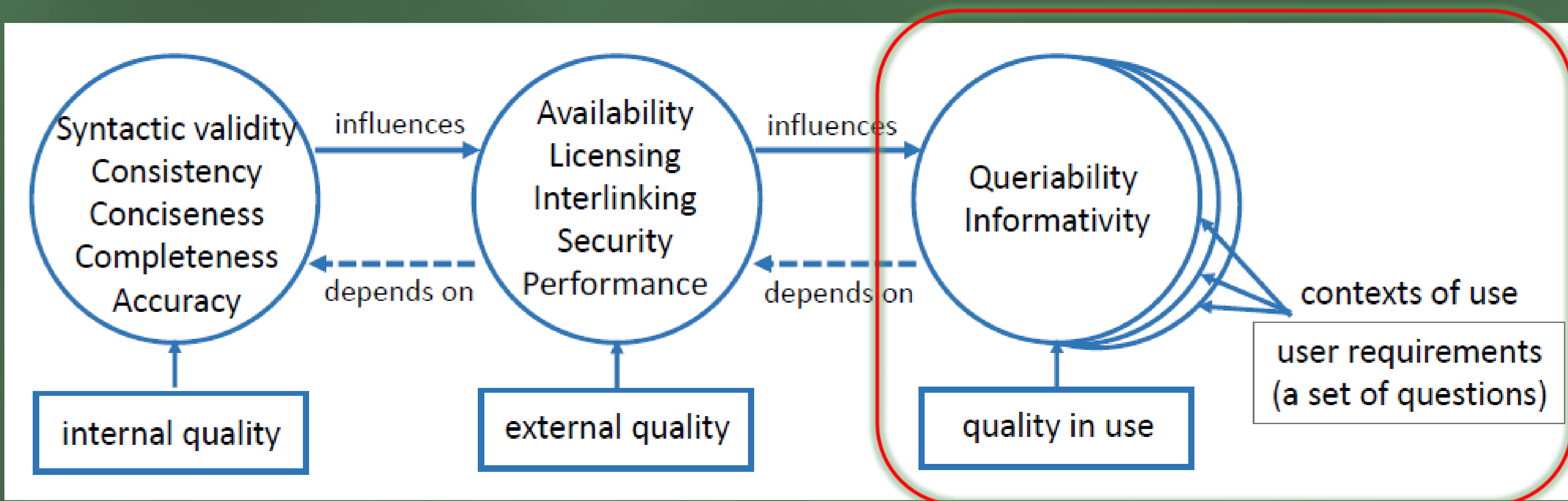
Brief Summary on the Experiment Result

- DBpedia Chinese and English are complementary to some extent. We find the person domain are more language-specific than place with the help of our comparative metrics.
- The correctness ratio of DBpedia Chinese and DBpedia English are almost the same statistically.
- Zhishi.me and SSCO have more instances than DBpedia Chinese, since they consolidate data from three Encyclopedic Web sites.
- The correctness ratio of YAGO from our assessment is not as high as declared, and we give a detail analysis of the reason for the phenomenon.
- The metrics on Overlapped Instances are not only useful in investigating the real quality of each KB, but also helpful to trace the characteristic of each KB.
- Multiple data sources may increase richness of data. However, whether it could improve the correctness of the KB depends on how we utilize the data sources.

Some recent results

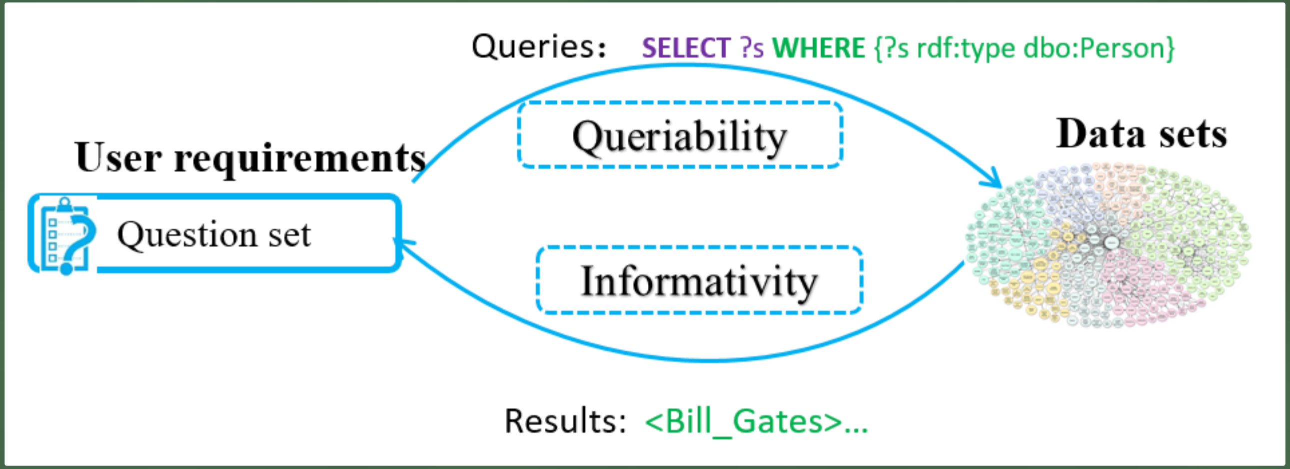
- Data quality assessment standard in big data transaction
- Data quality assessment on internal quality
- Data quality assessment on quality in use
- Data quality assessment on EMR (on-going)

Evaluation on quality in use



Ruan T, Li Y, Wang H, et al. From Queriability to Informativity, Assessing "Quality in Use" of DBpedia and YAGO[C]//International Semantic Web Conference. Springer International Publishing, 2016: 52-68.
Ruan T, Li Y, Wang H, et al. From Queriability to Informativity, Assessing "Quality in Use" of DBpedia and YAGO[C]//International Semantic Web Conference. Springer International Publishing, 2016: 52-68.

“Quality in use” are based on “Context of use”

- Data quality is fitness for use in special application context
- "context of use" refers as usage context or usage scenario, and it implies user requirements
- We use question sets to represent user requirements
- 

- Queriability measures how easily an end user can construct a correct query on a dataset.
- Informativity shows how informative a data set is under a particular usage context

Queriability Metrics

We design two kinds of Queriability metrics: subjective and (how easily objective

- Subjective metrics
  - Difficulty Rating, after evaluators finish constructing a query, they give ratings on how difficulty the process was. The rating has five levels: (1) very easy, (2) easy, (3) average, (4) difficult, and (5) very difficult.
- Objective metrics
  - Query Construction Time *(T)*
  - Query Construction Time On Domain *(T_a)*
  - Query Construction Time On Property Constraint *(T_b)*
  - Number of Attempts *(NOA)*

Informativity Metrics

- Subjective metrics
  - Informativeness Rating, is the users' rating on the informativeness that the results contain. It has five levels:

    (1) very little information (2) little information
    (3) some information (4) a fair amount of information
    (5) lots of information.
- Objective metrics
  - Precision
  - Recall
  - Comprehensive Informativity (*CI*)

Comprehensive Informativity (CI)

- Comprehensive Informativity (CI) is a metric which integrates different factors that influence the users' comprehension of information.
- Influence Factors
  - Precision
  - Recall
  - Accuracy of datasets (α)
  - Understandability of datasets (β)

$$CI = \frac{NCA}{NA} * (\frac{NCA}{A})^2 * \alpha * \beta$$

- NCA/NA is recall, NCA/A is precision. The square function is used to punish the irrelevant answers

Evaluated Knowledge Bases

- DBpedia2014 and YAGO2S

KBs	#class	#property	#instance	#fact
DBpedia	683	451k	4.58 M	68.1 M
YAGO	56k	75	2.89 M	7.8 M

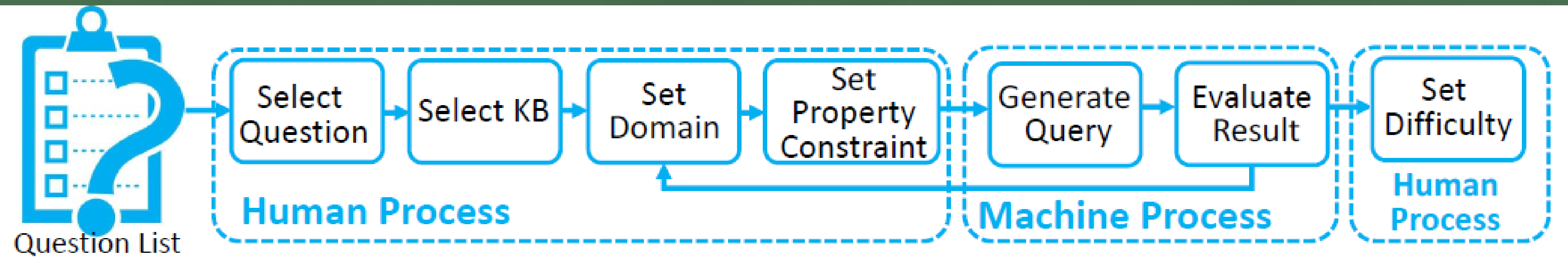
- The class in YAGO is much richer than DBpedia
- The number of property and fact are very huge

Questions

- Data requirements of users are modeled as question sets
- Two sources
  - QALD[1] (Question Answering over Linked Data)
  - WebQuestion[2] from the NLP laboratory of Standford
- 13 questions from QALD and 13 from WebQuestion
  - Give me all female Russian astronauts.
  - Which countries have more than two official languages?...

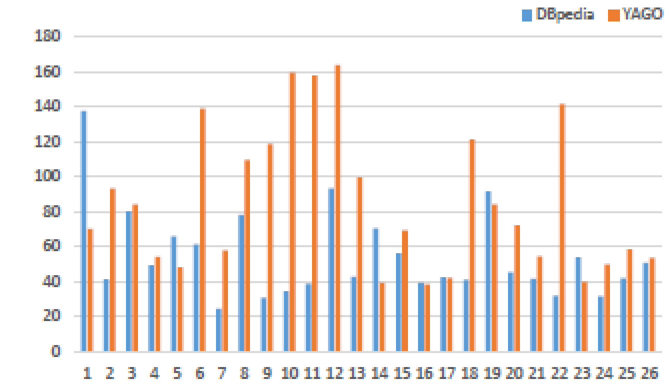
1. http://greententacle.techfak.uni-bielefeld.de/cunger/qald/index.php?x=home&q=home
2. http://nlp.stanford.edu/software/sempre/

Evaluation process

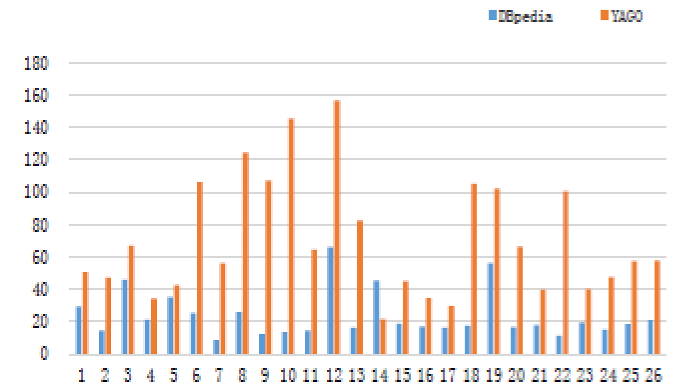


- Each evaluator first chooses a question from the question list and selects a target KB (YAGO or DBpedia).
- Then sets the domain of the question and inputs property constraints to construct a corresponding query
- Executable query is generated and run in the KB
- The results are returned to the evaluator, and the evaluator give subject evaluation

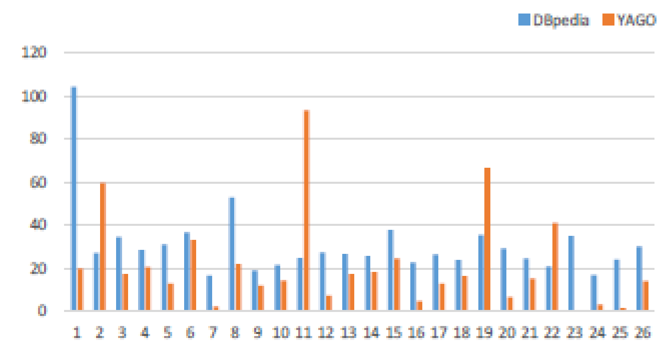
It costs evaluators more time on YAGO than on DBpedia to find a satisfactory query



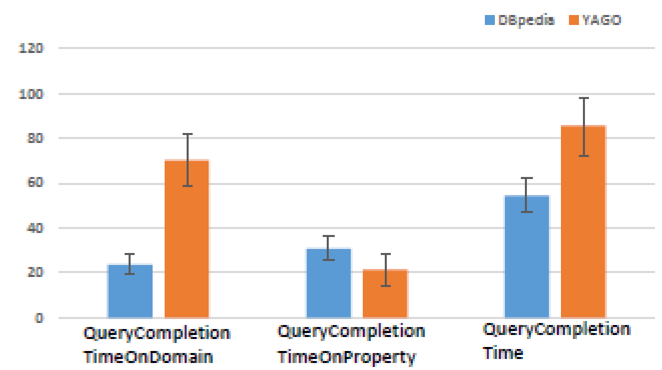
(a) The average Query Completion Time



(b) The average Query Completion Time On Domain



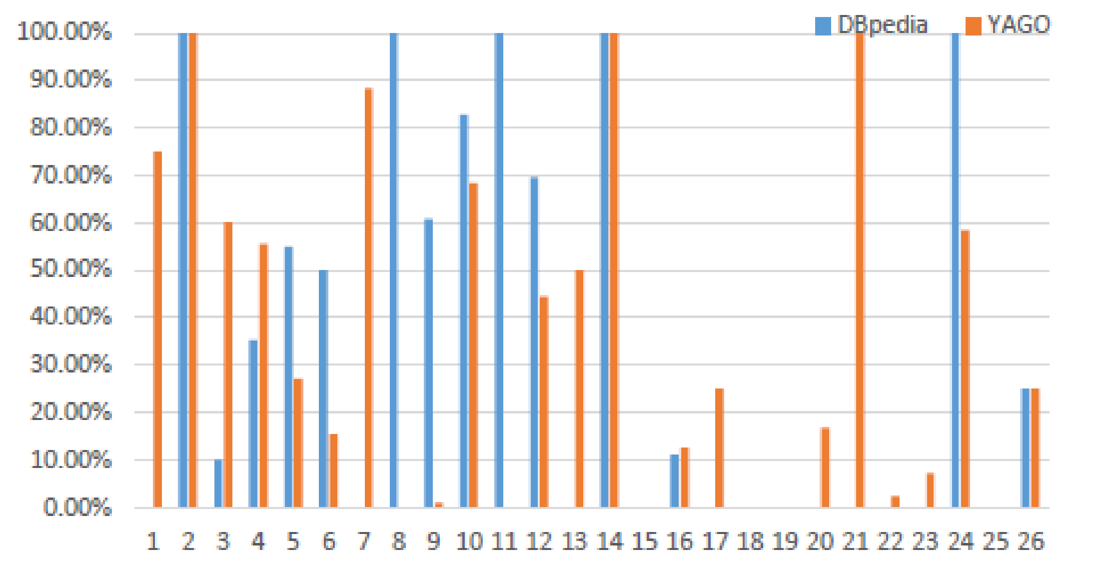
(c) The average Query Completion Time On Property Constraint



(d) The mean of Query Completion Time

- When users select domains, they spend much more time on YAGO than on DBpedia
- When considering properties selection, it takes longer on DBpedia than on YAGO

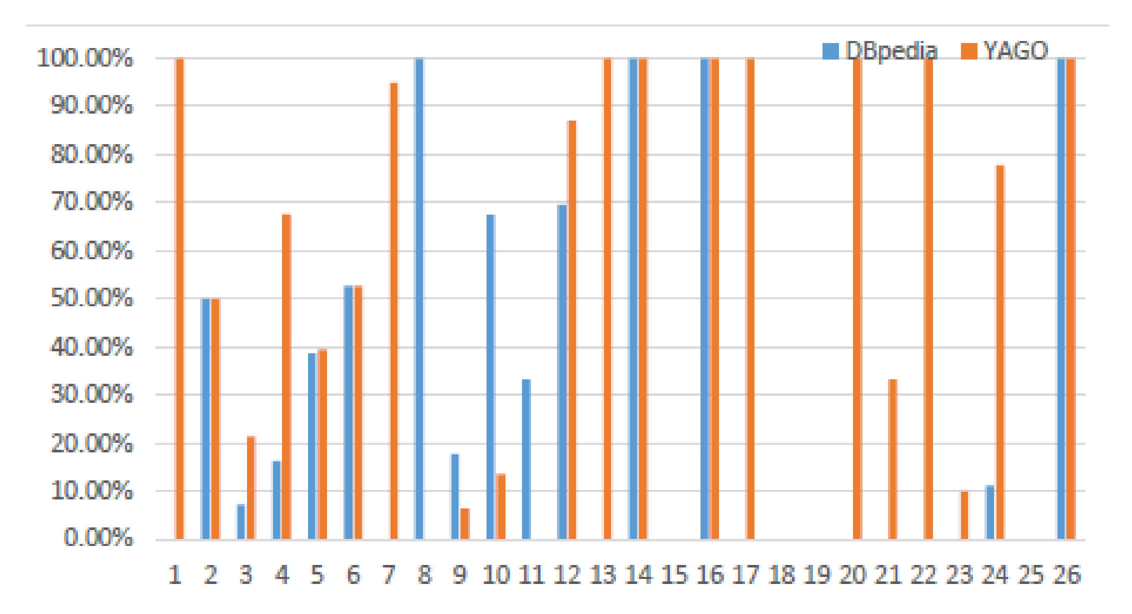
DBpedia has a rather low precision



(a) Precision

- DBpedia: the polysemy of type words. For example, presidents can be presidents of countries or presidents of organizations
- DBpedia example: “Give me all presidents of the United States,”, the results include the Chairman of the Federal Reserve and other types of presidents
- YAGO Example: the class “wikicategory German actors” means actors whose nationality, not birthplace, is German

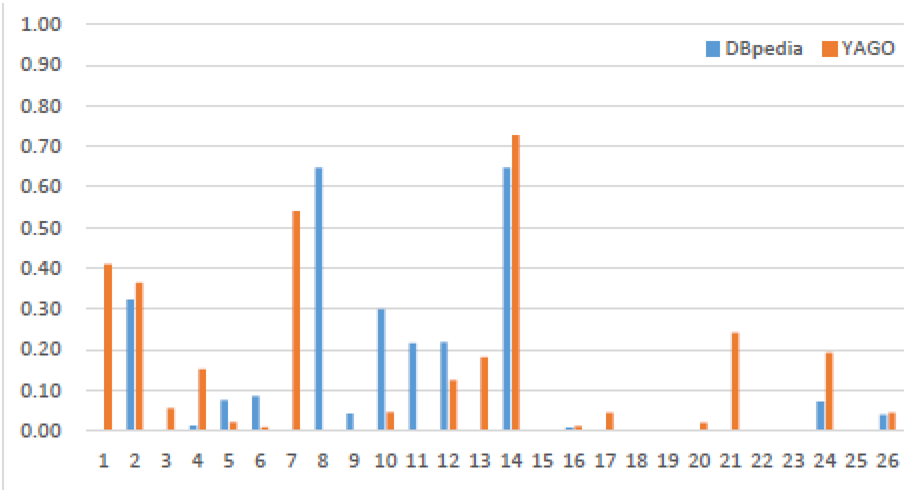
DBpedia and YAGO have low recalls



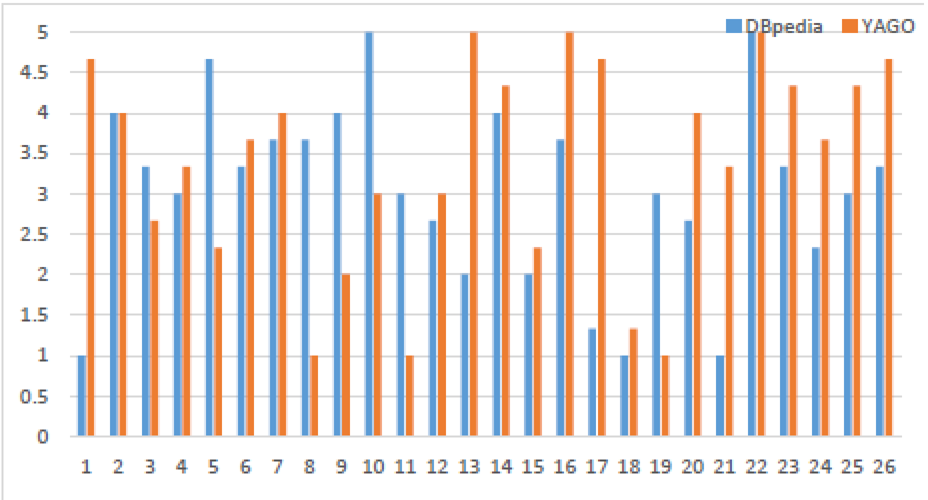
(b) Recall

- DBpedia: Duplicated properties as mentioned before.
- YAGO: it really does not contain abundant properties with enough facts

YAGO has slightly better results than DBpedia



(c) Comprehensive Informativity



(d) Informativeness Rating

- In summary, the results in DBpedia and YAGO both are not satisfied, and YAGO is slightly better than DBpedia

Assessment results in general

- Difficult to construct a query
- Low precision
- Low recall

lessons learned from our assessment work—— if datasets are intended to be used by human instead of computers

- Naming convention for classes, objects and properties are required in the LOD world, similar to that in software
- Duplicate property names should be avoided since they will mislead the users
- There should be a tradeoff between the number of classes and the number of properties.

Some recent results

- Data quality assessment standard in big data transaction
- Data quality assessment on internal quality
- Data quality assessment on quality in use
- Data quality assessment on EMR (on-going)

Examples of quality problems in EMR (1)

- Electronic Health Record data problems found in a hospital in Shanghai
- 1. Non-uniform
  - Sex of patients: 男，女，男性，女性
  - Age of patients: 85Y，65岁，55
- 2. Inconsistency

disease_class	disease_code	disease_name
ICD10	E14.900	2型糖尿病
ICD10	E14.901	2型糖尿病 糖尿病肾病
ICD10	E14.203+	2型糖尿病 糖尿病肾病
ICD10	E14.901	2型糖尿病 糖尿病肾病
ICD10	E14.901	2型糖尿病 糖尿病肾病
ICD10	E11.901	2型糖尿病 糖尿病肾病

Examples of quality problems (2)

- 3. Incomplete

	name	reference	rowkey	unit	groupid	serial_no
37		<60		pg/mL	3	2,010,231
02		1.84-7.2			8	2,010,305
02		3.13-6.42			9	2,010,306
33		0.5-5%			16	2,010,439
33		0.05-0.5 10^9/L			17	2,010,440
57		2.0-7.0		10^9/L	9	2,010,512
57		50-70		%	10	2,010,513
01		3.4-17umol/L			17	2,010,589
01		200-400		mg/L	18	2,010,590
01		20-35		g/L	19	2,010,591
01		15-41		IU/L	20	2,010,592
01		155-277mg/L			21	2,010,593
19		20-35		g/L	35	2,010,664
19		15-41		IU/L	36	2,010,665
20		250-400			30	2,010,735

- 4. InCorrect

Doctors input ICD code casually

Assessment dimensions on EMR data

Macro

- Breadth: measures a patient's record contains all desired types of data
- Density: measures a patient's record contains a specified number or frequency of data points over time
- Predictive: measures a patient's record contains sufficient information to predict a phenomenon of interest

Micro

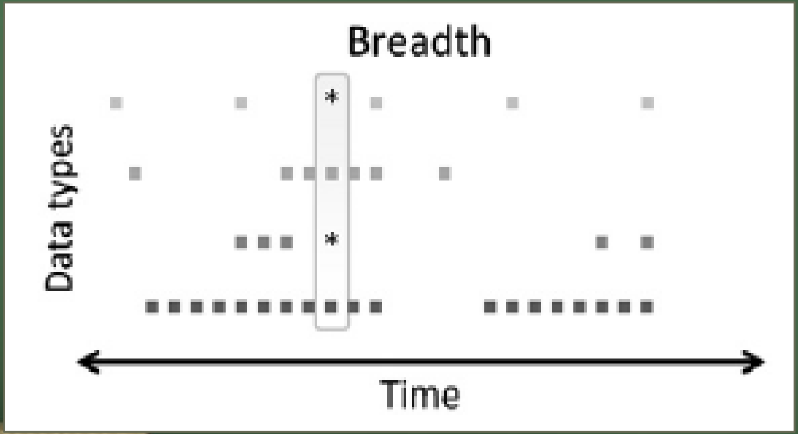
- Completeness:
- Correctness:
- Consistency:
- Normalization: measures normalization of data representation

Examples of assessment metrics on EMR data

- Completeness of patients' age
  - $\frac{\#.\,patient\ having\ no\ age}{\#.\,patient}$
- Correctness of patients' age
  - $\frac{\#.\,patient\ having\ correct\ age}{\#.\,patient\ having\ age}$
  - The true age can be calculated according to the patient's ID and In-patient time
- Consistency between patient's disease code and disease name
  - $\frac{\#.\,patient\ having\ consistent\ disease\ code\ and\ name}{\#.\,patient\ having\ both\ disease\ code\ and\ name}$
- Normalization of patients' age
  - $\frac{\#.\,patient\ having\ valid\ representation\ of\ age}{\#.\,patient\ having\ age}$

Macro Metrics — Breadth

- Some secondary use scenarios require the availability of multiple types of data (disease prediction, risk analysis)
- For example, diagnoses, laboratory results, medications, and procedure codes
- $\frac{\#.\ patient\ having\ all\ above\ data\ items}{\#.\ patient}$



Weiskopf N G, Hripcsak G, Swaminathan S, et al. Defining and measuring completeness of electronic health records for secondary use[J]. Journal of biomedical informatics, 2013, 46(5): 830-836.

Macro Metrics — Density

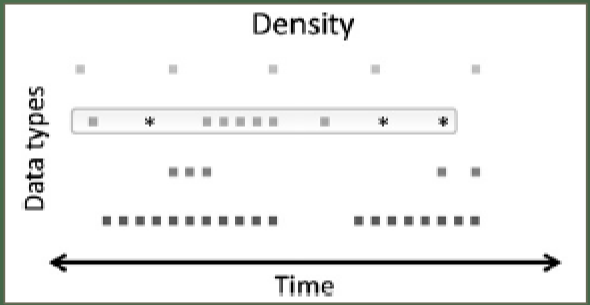
- EHR data consumers require not only a breadth of data types, but also sufficient numbers and density of data points over time
- Suppose, for a patient had laboratory examination n times, the examination times are x1, x2, …, xn.

$$I = \frac{2}{n} + \frac{n-2}{n}\left[1 - \sqrt{(n-1)\mathrm{Var}\{g_i; i = 1, \dots, n-1\}}\right].$$

average amount of information that each laboratory result examines

$$g_i = \frac{x_{(i+1)} - x_{(i)}}{x_{(n)} - x_{(1)}}$$

relative time gaps between examination

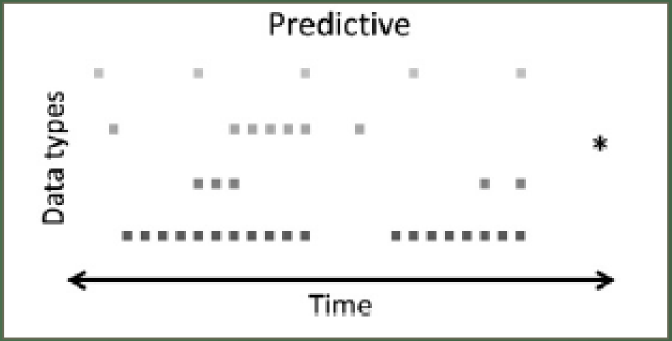


- $n_e = I \times n$ is the effective number of laboratory results available for each patient, namely, the density of laboratory results for each patient

Weiskopf N G, Hripcsak G, Swaminathan S, et al. Defining and measuring completeness of electronic health records for secondary use[J]. Journal of biomedical informatics, 2013, 46(5): 830-836.

Macro Metrics — Predictive

- The overall goal of much research on EHR is the ability to predict an outcome (disease status, risk, readmission, mortality)



- $\frac{\#.\, patient\ with\ correct\ prediction\ outcome}{\#.\, patient}$

Weiskopf N G, Hripcsak G, Swaminathan S, et al. Defining and measuring completeness of electronic health records for secondary use[J]. Journal of biomedical informatics, 2013, 46(5): 830-836.

Reference

1. Fürber C. Data Quality in the Semantic Web[M]//Data Quality Management with Semantic Technologies. Springer Fachmedien Wiesbaden, 2016: 69-77.
2. Strong, D.M., Lee, Y.W., Wang, R.Y.: Data quality in context. Communications of the ACM 40(5) (1997) 103-110
3. Fan W, Geerts F. Foundations of data quality management[J]. Synthesis Lectures on Data Management, 2012, 4(5): 1-217.
4. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: A survey. Semantic Web Journal 7 (2016) 63-93
5. Christian Bizer and Richard Cyganiak. Quality-driven information filtering using the wiqa policy framework. Journal of Web Semantics, 7(1):1–10, 2009.
6. Christian Fürber and Martin Hepp. SWIQA - a semantic web information quality assessment framework. In ECIS, 2011.
7. Aidan Hogan, Andreas Harth, Alexandre Passant, Stefan Decker, and Axel Polleres. Weaving the pedantic web. In 3rd Linked Data on the Web Workshop at WWW, 2010.
8. Yuangui Lei, Victoria Uren, and Enrico Motta. A framework for evaluating semantic metadata. In 4th International Conference on Knowledge Capture, number 8 in K-CAP ’07, pages 135 – 142. ACM, 2007.
9. Jens Lehmann, Daniel Gerber, Mohamed Morsey, and Axel-Cyrille Ngonga Ngomo. DeFacto - Deep Fact Validation. In ISWC. Springer Berlin / Heidelberg, 2012.
10. Yang Yu and Jeff Heflin. Extending functional dependency to detect abnormal data in rdf graphs. In ISWC, volume 7031, pages 794–809. Springer Berlin Heidelberg, 2011.
11. Flemming, Annika Quality characteristics of linked data publishing datasources. MA thesis. Humboldt-Universität zu Berlin, Institut für Informatik, 2011.
12. Mendes, P.N., M□uhleisen, H., Bizer, C.: Sieve: linked data quality assessment and fusion. In: Proceedings of the 2012 Joint EDBT/ICDT Workshops, ACM (2012) 116-123.
13. A. Schultz, A. Matteini, R. Isele, C. Bizer, and C. Becker. Ldif : Linked data integration framework. 2011.
14. Kontokostas, D., Zaveri, A., Auer, S., Lehmann, J.: Triplecheckmate: A tool for crowdsourcing the quality assessment of linked data. Knowledge Engineering and Semantic Web (2013) 265-272
15. Feeney K C, O'Sullivan D, Tai W, et al. Improving curated web-data quality with structured harvesting and assessment[J]. International Journal on Semantic Web and Information Systems (IJSWIS), 2014, 10(2): 35-62.
16. ISO/IEC25010: Systems and software engineering -Systems and software Quality Requirements and Evaluation (SQuaRE)-System and software quality models.(2011)
17. Tong Ruan, Xu Dong, Haofen Wang, and Yang Li. Evaluating and Comparing Web-Scale Extracted Knowledge Bases in Chinese and English. JIST2015
18. Ruan T, Li Y, Wang H, et al. From Queriability to Informativity, Assessing “Quality in Use” of DBpedia and YAGO[C]//International Semantic Web Conference. Springer International Publishing, 2016: 52-68.
19. Weiskopf N G, Hripcsak G, Swaminathan S, et al. Defining and measuring completeness of electronic health records for secondary use[J]. Journal of biomedical informatics, 2013, 46(5): 830-836.
20. Puttkammer N, Baseman J G, Devine E B, et al. An assessment of data quality in a multi-site electronic medical record system in Haiti[J]. International journal of medical informatics, 2016, 86: 104-116.

Thank you

Q&A