# Knowledge Graphs

Dieter Fensel et al.

Chapter 2.4 ~ 2.5

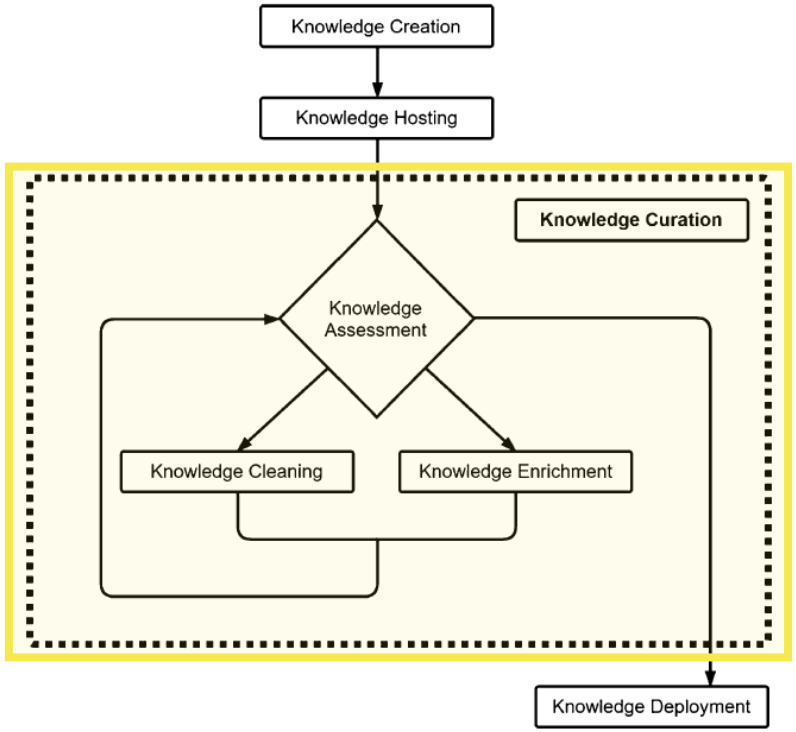
SoYeop Yoo

# CONTENTS

# 2.4 KNOWLEDGE CURATION

## Knowledge Curation

- Overall goal
  - To provide cost-sensitive methods to improve the quality of large knowledge graphs ensuring their usefulness for envisaged applications



# 2.4 KNOWLEDGE CURATION

## A Maximal Simple Knowledge Representation Formalism

- Representation formalism
  - In sync with schema.org
  - TBox
    - Schema definitions
    - Define types, hierarchy, properties with their domains and ranges

    - Two disjoint and finite sets of type and property names $T$ and $P$
    - A finite number of type definitions
      → $isA(t_1, t_2)$
    - A finite number of property definitions
      → $hasDomain(p, t)$
      → $hasRange(p, t_2)$
      → $hasLocalRange(p, t_1, t_2)$

# 2.4 KNOWLEDGE CURATION

## A Maximal Simple Knowledge Representation Formalism

- Representation formalism
  - ABox
    - Assertional statement
    - Add assertions over the terminology
    - A countable set of instance identifiers $I$
    - Instance assertions
      $\rightarrow isElementOf(i, t)$
    - Property value assertions
      $\rightarrow p(i_1, i_2)$
    - Equality assertions
      $\rightarrow isSameAs(i_1, i_2)$

# 2.4 KNOWLEDGE CURATION

- Knowledge Assessment
  - Describe and define the process of assessing the quality of a KG
  - Goal
    - To measure the usefulness of a KG considering correctness and completeness

# 2.4 KNOWLEDGE CURATION

## Knowledge Assessment

### Literature

- Accessibility
- Accuracy (veracity)
- Appropriate amount (relevancy)
- Believability (trustworthiness)
- Completeness
- Concise representation
- Consistent representation
- Cost-effectiveness
- Ease of manipulation
- Ease of operation
- Ease of understanding
- Flexibility
- Free-of-error (accuracy)
- Interoperability
- Objectivity (part of Trustworthiness)
- Relevancy
- Reputation (part of Trustworthiness)
- Security (part of Accessibility)
- Timeliness (velocity)
- Traceability (verifiability)
- Understandability (ease of understanding)
- Value-added
- Variety

# 2.4 KNOWLEDGE CURATION

## Knowledge Assessment

### Task Types

- Given maximal simple knowledge representation formalism → distinguish three error sources
  - Instance assertions: $isElementOf(i, t)$
  - Property value assertions: $p(i_1, i_2)$
  - Equality assertions: $isSameAs(i_1, i_2)$
- Considering the correctness and completeness → define six tasks
  - Correctness
    - →Identify the number of wrong instance assertions
    - →Identify the number of wrong property value assertions
    - →Identify the number of wrong equality assertions
  - Completeness
    - →Identify the number of missing instance assertions
    - →Identify the number of missing property value assertions
    - →Identify the number of missing equality assertions

# 2.4 KNOWLEDGE CURATION

## Knowledge Assessment

### Methods and Tools

#### Methodologies

- Total Data Quality Management (TDQM) (Wang 1998)
- Data Quality Assessment (Pipino et al. 2002)
- User-driven assessment (Zaveri et al. 2013)
- Test-driven assessment (Kontokostas et al. 2014)
- Manual assessment based on human expertise (Acosta et al. 2013)
- Statistical distribution for measuring the correctness of statements (Paulheim and Bizer 2014)

#### Tools

- WIQL (Web Information Quality Assessment Framework)
- SWIQA (Semantic Web Information Quality Assessment Framework)
- LINK-QA
- Sieve
- Validata
- Luzzu

# 2.4 KNOWLEDGE CURATION

## Knowledge Assessment

### Methods and Tools

- Approaches that help to define metrics and tackle the assessment of the tasks
  - Sieve for Data Quality Assessment (Mendes et al. 2012)
    - →Data quality indicators
    - →Scoring functions
    - →Assessment metrics
    - →Aggregate metrics
  - Using Semantic Web Resources for Data Quality Management (Furber and Hepp 2010)
    - →Missing literal values
    - →False literal values
    - →Functional dependency violations
  - SDType (Paulheim and Bizer 2013)
    - →Statistical distribution
  - Resolving Range Violations (Lertvittayakumjorn et al. 2017)
    - →Reduce search space
    - →Calculating scores

# 2.4 KNOWLEDGE CURATION

## Knowledge Cleaning

- Goal
  - To improve the correctness of a KG
- Major objectives
  - Error detection
    - Identifying wrong assertions in a KG
  - Error correction
    - Correcting wrong assertions by deleting or modifying

# 2.4 KNOWLEDGE CURATION

## Knowledge Cleaning

### Literature

- Most of the approaches
  - Focus on the error detection or correction → not for both tasks
  - Focus on detecting errors than correcting
- Error detection
  - SDType (Paulheim and Bizer 2013)
  - SDValidate (Paulheim and Bizer 2014)
- Error correction
  - Katara (Chu et al. 2015)
    - →Correcting wrong instance and property value assertions
  - HoloClean (Rekatsinas et al. 2017)
    - →Correcting wrong property value assertions
  - LOD Laundromat (Beek et al. 2014)
    - →Correcting syntax errors

# 2.4 KNOWLEDGE CURATION

## Knowledge Cleaning

### Task Types

- Detection/Correction of wrong instance assertions: $isElementOf(i, t)$
  - → $i$ is not a proper instance identifier
  - → $t$ is not an existing type name
  - → The instance assertion is wrong
- Detection /Correction of wrong property value assertions: $p(i_1, i_2)$
  - → $p$ is not a proper property name
  - → $i_1$ is not a proper instance identifier
  - → $i_1$ is not in any domain of $p$
  - → $i_2$ is not a proper instance identifier
  - → $i_2$ is not in any ranger of $p$
  - → The property assertion is wrong
- Detection /Correction of wrong equality assertions: $isSameAs(i_1, i_2)$
  - → $i_1$ is not a proper instance identifier
  - → $i_2$ is not a proper instance identifier
  - → The identity assertion is wrong

# 2.4 KNOWLEDGE CURATION

## Knowledge Cleaning

### Methods and Tools

#### Methods according to the cleaning target

- Instance assertion
  - →A statistical distribution of types and properties
  - →Supervised machine learning and entity-type dictionaries
  - →Association rule mining
- Property value assertion
  - →Statistical distribution
  - →Ontology reasoners
  - →Wikipedia pages
  - →Outlier detection
- Equality assertion
  - →Outlier detection
  - →Constraints
  - →Logical verification
  - →Local context of instances

# 2.4 KNOWLEDGE CURATION

- Knowledge Cleaning
  - Methods and Tools
    - Tools
      - HoloClean (Rekatsinas et al. 2017)
      - KATARA (Chu et al. 2015)
      - SDValidate (Paulheim and Bizer 2014)
      - SPIN (SPARQL Inferencing Notation)
      - LOD Laundromat (Beek et al. 2014)
      - TISCO (Temporal Scoping of Facts) (Rula et al. 2019)

# 2.4 KNOWLEDGE CURATION

## Knowledge Enrichment

- Goal
  - To improve the completeness of a KG by adding new statements
- Process
  - Identify a relevant knowledge source
  - Integrate of TBox and ABox statements from the source to our KG
    - Issues
      - →Merging or aligning different schemata (TBox)
      - →Identifying and resolving duplicates (ABox)
      - →Invalid property statements (ABox)

# 2.4 KNOWLEDGE CURATION

## Knowledge Enrichment

### Literature

- Knowledge management shares some goals with data management
  - Data access, data quality, data cleansing, data integration, and more
  - Knowledge enrichment = data fusion
- Issues to produce more consistent, accurate, and useful knowledge
  - Entity resolution
    - →Deriving new $isSameAS(instance_1, instance_2)$ assertions and aligning the descriptions of these two identifiers
  - Resolving conflicting property value assertion
    - →Handling for example situations such as $P(i_1, i_2)$, and $P(i_1, i_3)$, and $i_2 \neq i_3$, and $P$ has a unique value constraint
    - →Refer to error detection and correction

# 2.4 KNOWLEDGE CURATION

## Knowledge Enrichment

- Task Types
  - Identifying and resolving duplicates (lack of $isSameAs(i_1, i_2)$ assertions)
  - Resolving conflicting property value assertions

# 2.4 KNOWLEDGE CURATION

## Knowledge Enrichment

### Methods and Tools

#### Methods for identifying duplicates

- String similarity measures (Winkler 2006)
- Association rule mining (Hipp et al. 2000)
- Topic modeling (Sleeman et al. 2015)
- Support vector machine (Sleeman and Finin 2013)
- Property-based (Hogan et al. 2007)
- Crowd-sourced data (Getoor and Machanavajjhala 2013)
- Graph-oriented techniques (Korula and Lattanzi 2014)

# 2.4 KNOWLEDGE CURATION

## Knowledge Enrichment

### Methods and Tools

- Methods and tools for identifying and resolving duplicates
  - ADEL (Adaptable Entity Linking) (Plu et al. 2017)
    → Indexing and linking data based on the label and popularity of entities
  - Dedupe
    → Python library that uses machine learning to find duplicate entries
  - Duke (Garshol and Borge 2013)
  - Legato (Achichi et al. 2017)
  - LIMES (Ngomo and Auer 2011)
  - SERIMI (Araujo et al. 2011)
  - Silk (Volz et al. 2009)

# 2.4 KNOWLEDGE CURATION

- Knowledge Enrichment
  - Methods and Tools
    - Methods and tools for resolving conflicting property value assertions
      - FAGI (Giannopoulos et al. 2014)
      - FuSem (Bleiholder et al. 2007)
      - HumMer (Bilke et al. 2005)
      - KnoFuss (Nikolov et al. 2008)
      - ODCleanStore (Knap et al. 2012)
      - Sieve (Mendes et al. 2012)

# 2.5 KNOWLEDGE DEPLOYMENT

## Knowledge Deployment

- Linked Open Data (LOD)
  - Open data
    - Data that can be freely used, re-used and redistributed by anyone
  - Linked data
    - A method of publishing structured data so that it can be interlinked and become more useful through semantic queries
  - LOD = Open data + Linked data
    - Linked data published as open data or Open data published as linked data
  - 5* criteria of the quality of LOD
    - *) if the data are provided under an open license
    - **) if the data are available as structured data
    - ***) if the data are also available in a non-proprietary format
    - ****) if URIs are used so that the data can be referenced
    - *****) if the dataset is linked to other datasets to provide context

# 2.5 KNOWLEDGE DEPLOYMENT

## Knowledge Deployment

- Overview of some KGs

| Name | Instances | Facts | Types | Relations |
|---|---|---|---|---|
| DBpedia (English) | 4,806,150 | 176,043,129 | 735 | 2813 |
| YAGO | 4,595,906 | 25,946,870 | 488,469 | 77 |
| Freebase | 49,947,845 | 3,041,722,635 | 26,507 | 37,781 |
| Wikidata | 15,602,060 | 65,993,797 | 23,157 | 1673 |
| NELL | 2,006,896 | 432,845 | 285 | 425 |
| OpenCyc | 118,499 | 2,413,894 | 45,153 | 18,526 |
| Google's Knowledge Graph | 570,000,000 | 18,000,000,000 | 1500 | 35,000 |
| Google's Knowledge Vault | 45,000,000 | 271,000,000 | 1100 | 4469 |
| Yahoo! Knowledge Graph | 3,443,743 | 1,391,054,990 | 250 | 800 |

# 2.5 KNOWLEDGE DEPLOYMENT

## Knowledge Deployment

- Building, implementing, and curating KG
  - Time-consuming and costly activities
  - The average cost for one fact in a KG
    - $0.1 ~ $6 depending on the amount of mechanization
- Alternatives of cost model
  - The data consumer is paying for service
  - The data supplier is paying for service (Open data or linked open data)