Linked Data & Semantic Web (2021-2A)

Ontology Matching

Shenghui Wang

22 March 2022

UNIVERSITY OF TWENTE.

Goals of today’s lecture

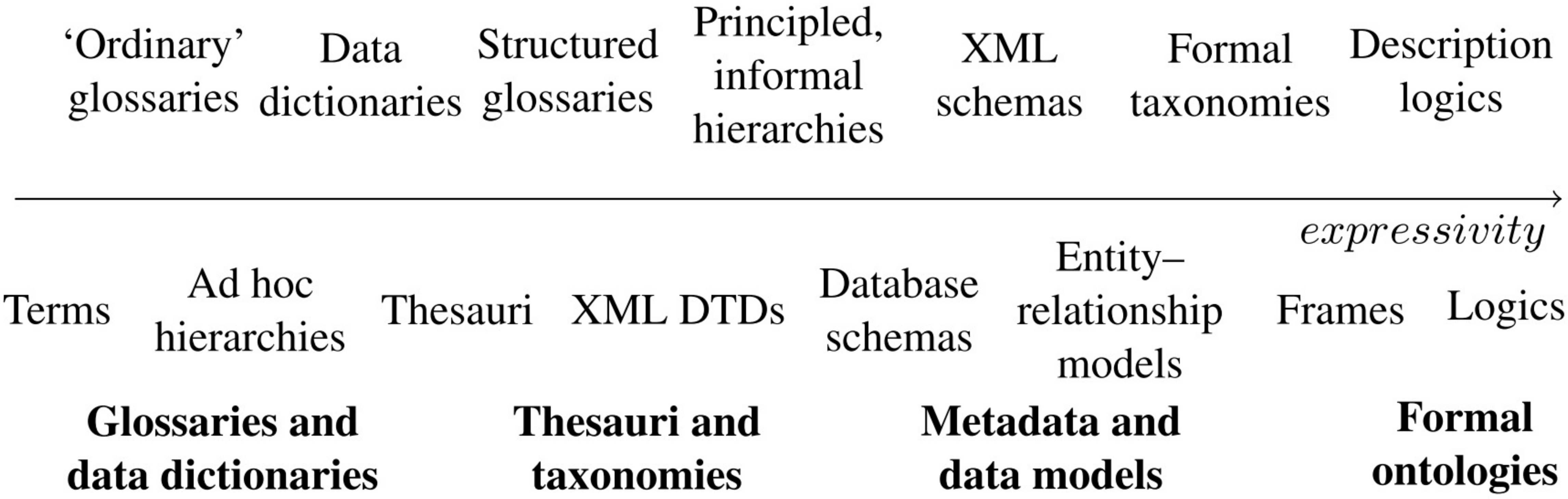
- Introduce what ontology matching is and why it is important
- Discuss practical and methodological issues
- Describe different techniques for ontology matching
- Evaluate the quality of ontology matching

Introduction

What is an ontology?

- An ontology typically provides a vocabulary describing a domain of interest and a specification of the meaning of terms in that vocabulary
- Depending on the precision of this specification, the notion of ontology encompasses several data and conceptual models, including, sets of terms, classifications, thesauri, database schemas, or fully axiomatized theories.

Various forms of ontologies



'Ordinary' glossaries

Data dictionaries

Structured glossaries

Principled, informal hierarchies

XML schemas

Formal taxonomies

Description logics

expressivity

Terms

Ad hoc hierarchies

Thesauri

XML DTDs

Database schemas

Entity–relationship models

Frames

Logics

Glossaries and data dictionaries

Thesauri and taxonomies

Metadata and data models

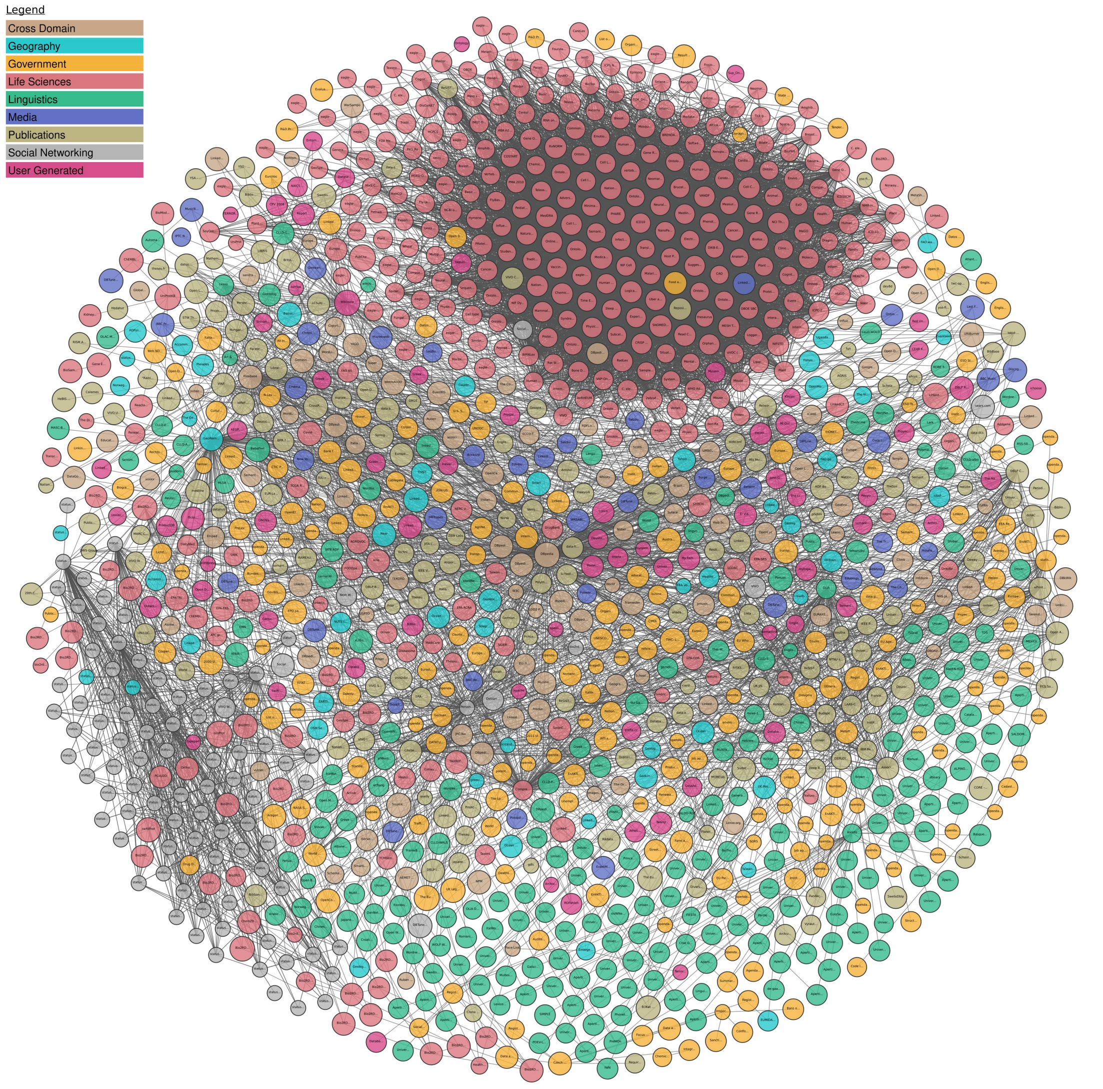
Formal ontologies

Euzenat, J. & Shvaiko, P. (2013). Ontology matching (2nd ed.). Springer Berlin Heidelberg.

Ontologies are not Reality

- Ontologies are a context-dependent projection (model) of the Reality
- Different ontologies might model the same (similar) or highly related domains, but they might
  - Reflect different tasks and requirements of applications
  - Follow different conventions and restrictions

Linked Open Data cloud



The Linked Open Data Cloud from lod-cloud.net

So it is a mess, but a meaningful one!

Legend

Cross Domain

Geography

Government

Life Sciences

Linguistics

Media

Publications

Social Networking

User Generated

https://lod-cloud.net/

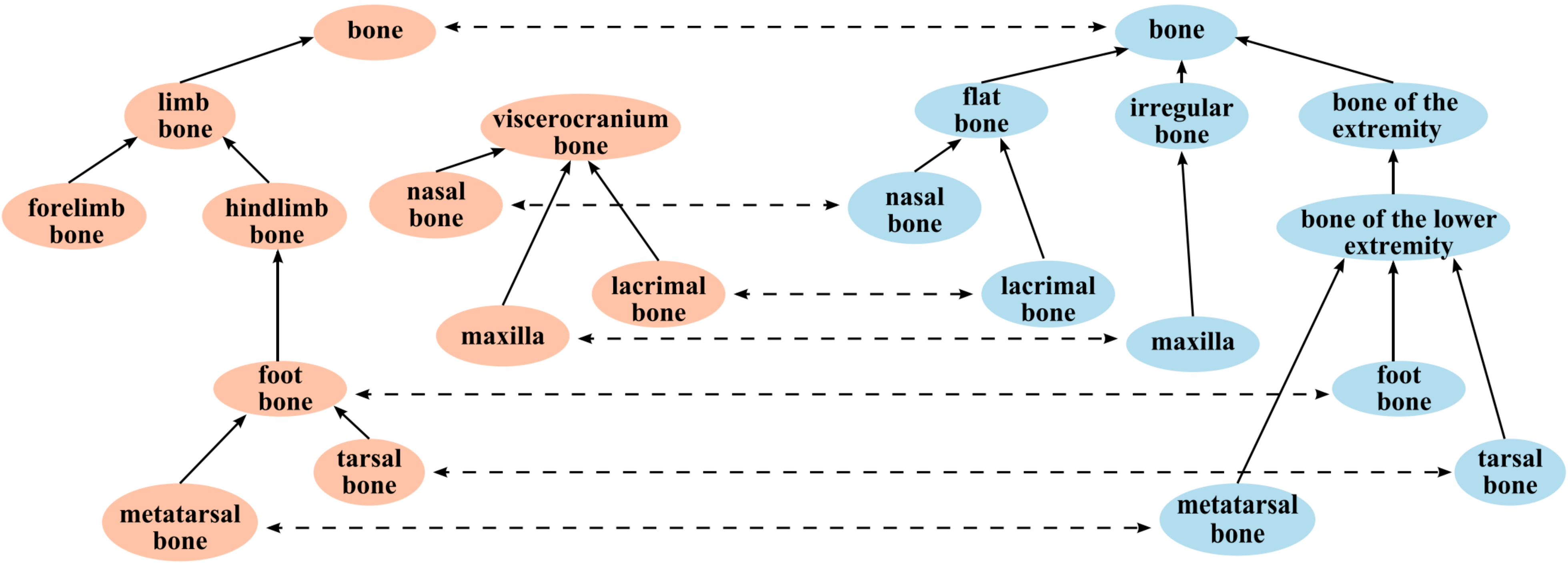
The heterogeneity problem

- Syntactic heterogeneity
  - Using different ontology languages (e.g. XML, OWL)
- Terminological heterogeneity
  - Different terms refer to the same concept
  - Same term describes different concepts
- Conceptual (Semantic) heterogeneity
  - Difference in *granularity, coverage, perspective*
- Semiotic (Pragmatic) heterogeneity
  - Different interpretations wrt different context

Ontology matching to reduce **terminological** & **conceptual** heterogeneity

What is ontology matching?

- Ontology matching, is the process of determining correspondences between concepts in ontologies.



Pambrix, Patrick & Ivanova, Valentina. (2013). A unified approach for debugging is-a structure and mappings in networked taxonomies. Journal of biomedical semantics. 4. 10. 10.1186/2041-1480-4-10.

Correspondence

- Given two ontologies o and o', a correspondence between o and o' is defined as $< id, e, e^{'}, r, n >$,

  where

  - id is a unique identifier of the correspondence
  - e and $e^{'}$ are entities of o and o' respectively, e.g., classes, instances
  - r is a relation, e.g., equivalence($=$), more general ($\sqsupseteq$), more specific ($\sqsubseteq$), disjointness ($\perp$), etc.
  - n is a confidence measure (typically in a range of [0,1]) for the correspondence between e and $e^{'}$.

Correspondence relations

- OWL
    - owl:equivalentClass, owl:equivalentProperty
    - owl:sameAs, owl:differentFrom, owl:AllDifferent
- SKOS
    - skos:exactMatch, skos:closeMatch
    - skos:broadMatch, skos:narrowMatch
    - skos:relatedMatch

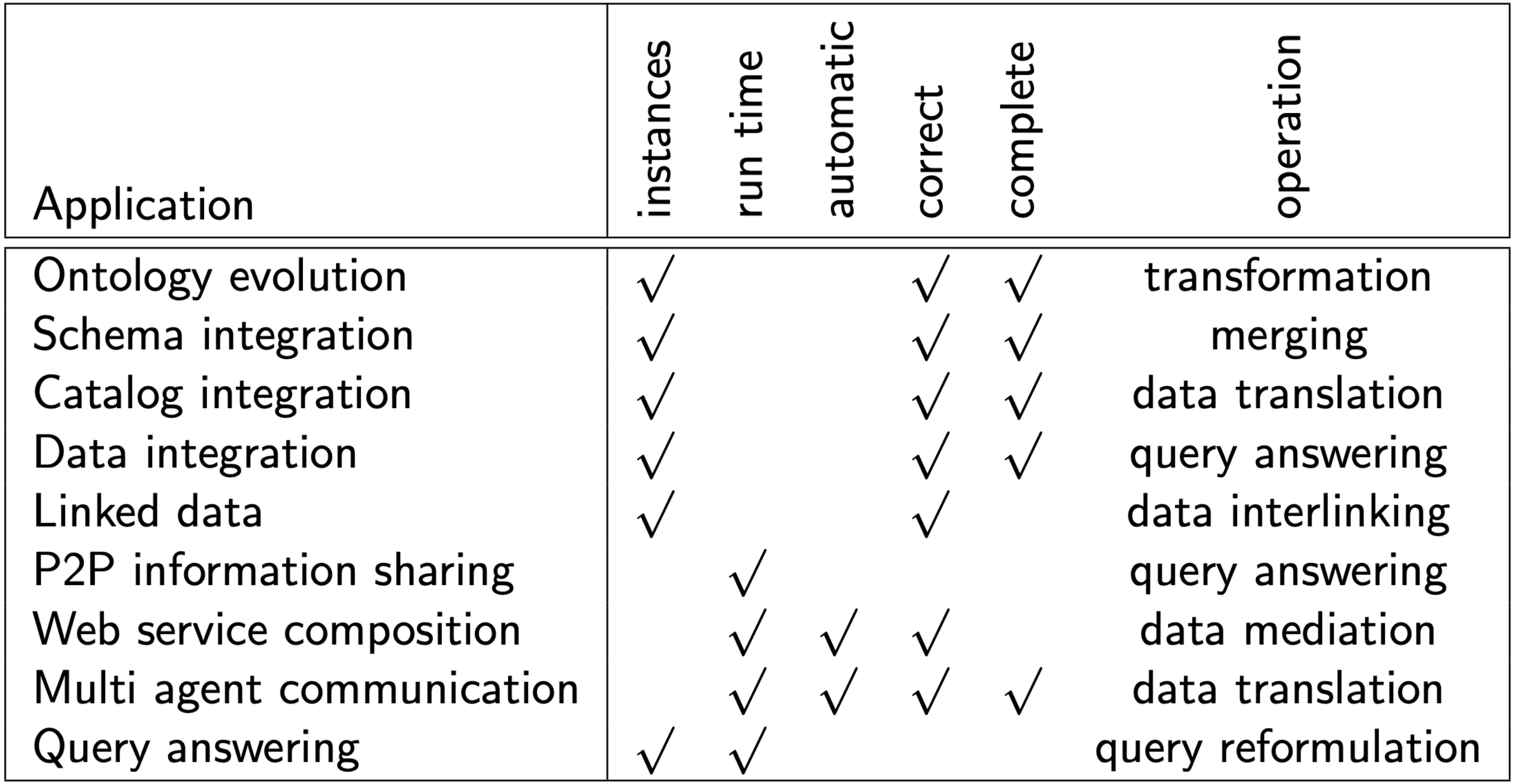
Alignment

- Given two ontologies o and o’, an alignment between o and o’:
  - is a set of correspondences on o and o’
  - with some additional metadata
    - multiplicity: 1:1, 1:m, m:n
    - method
    - date
    - creator
    - …

Terminology

- Matching is the process of finding relationships or correspondences between entities of different ontologies.
- Alignment is a set of correspondences between two or more (in case of multiple matching) ontologies. The alignment is the output of the matching process.
- Correspondence is the relation supposed to hold according to a particular matching algorithm or individual, between entities of different ontologies.
- Mapping is the oriented, or directed, version of an alignment.

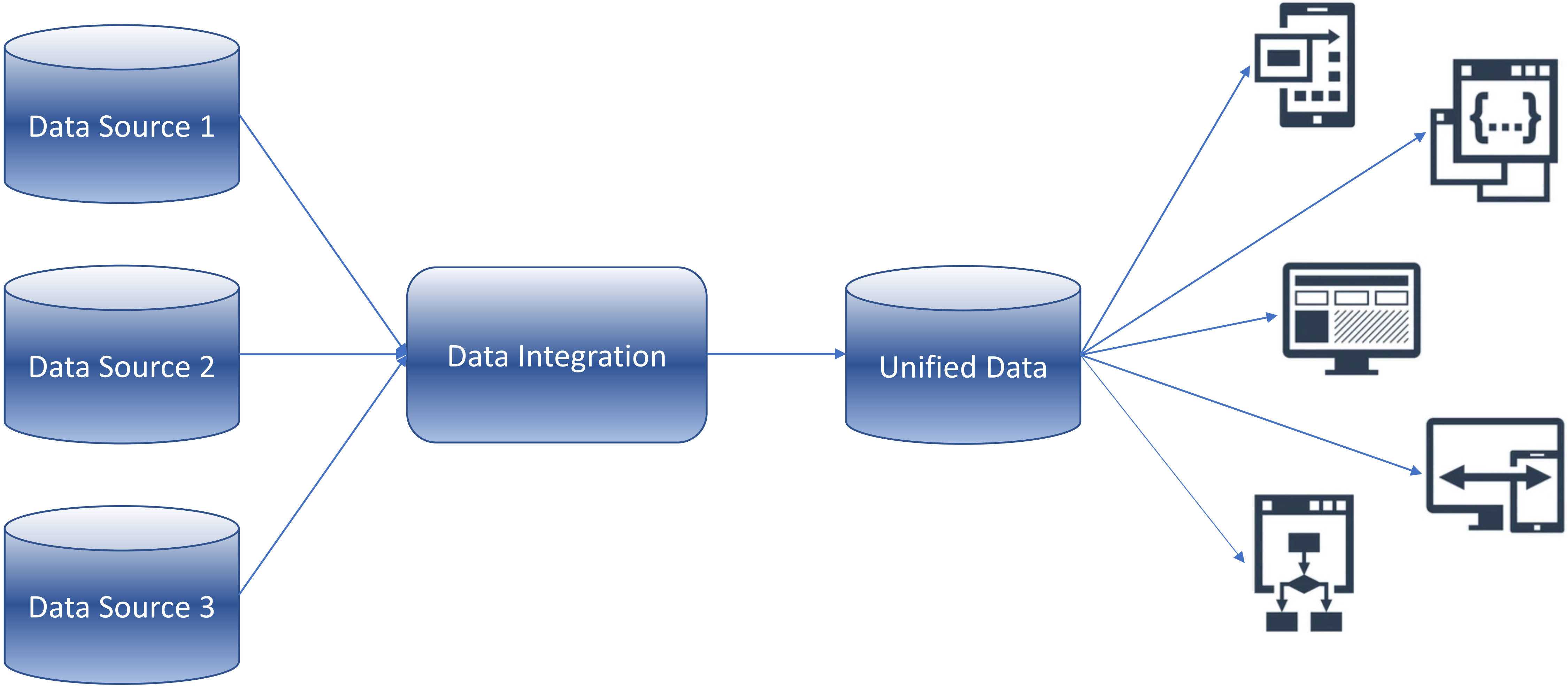
Ontology matching applications



Application	instances	run time	automatic	correct	complete	operation
Ontology evolution	√			√	√	transformation
Schema integration	√			√	√	merging
Catalog integration	√			√	√	data translation
Data integration	√			√	√	query answering
Linked data	√			√		data interlinking
P2P information sharing		√				query answering
Web service composition		√	√	√		data mediation
Multi agent communication		√	√	√	√	data translation
Query answering	√	√				query reformulation

http://tutorial.ontologymatching.org/

Data Integration



Data Source 1

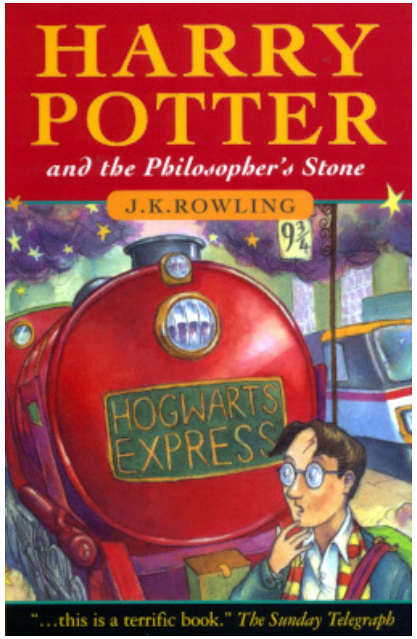
Data Source 2

Data Source 3

Data Integration

Unified Data

Let’s look at an example


HARRY POTTER
and the Philosopher's Stone
J.K.ROWLING
9¾
HOGWARTS EXPRESS
"...this is a terrific book." The Sunday Telegraph

- **Title**: Harry Potter and the Philosopher’s Stone
- **Author**: J. K. Rowling
- **Illustrator**: Thomas Taylor
- **Publication year**: 1997
- **Publisher**: Bloomsbury (UK)
- **ISBN**: 0-7475-3269-9

WIKIPEDIA
The Free Encyclopedia

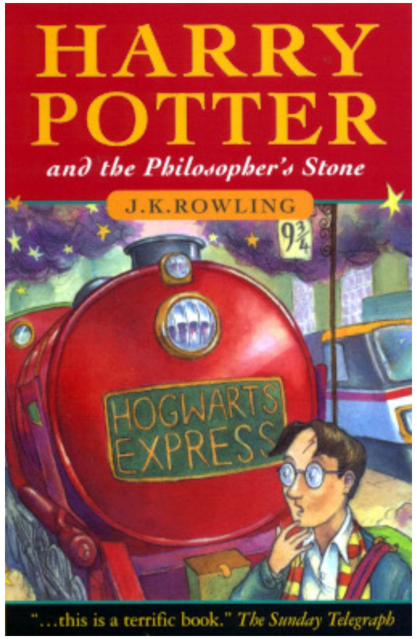
BRITISH LIBRARY

- **Title**: Harry Potter and the Philosopher’s Stone
- **Author**: J. K. Rowling
- **Subjects:** Juvenile fiction, fantasy fiction
- **Dewey**: 823.914 <- English fiction, between 1945 and 1999
- **Publication details**: London : Bloomsbury, 1997 2004 printing.
- **ISBN**: 0747574472 (pbk)

LIBRARY
LIBRARY OF CONGRESS

- **Title**: Harry Potter and the sorcerer’s stone
- **Uniform title**: Harry Potter and the philosopher’s stone
- **Personal name**: Rowling, J. K., author
- **Published/Produced**: New York, NY : Scholastic Inc., [2018]
- **ISBN**: 9781338299144 (paperback), 133829914X (paperback)
- **LC Subjects**: Wizards—Fiction, Magic-–Fiction, Schools-Fiction
- **Dewey class no.:** 823/.914 [Fic]
- **Summary**: Rescued from the outrageous neglect of his aunt and uncle, a young boy with a great destiny proves his worth while attending Hogwarts School for Witchcraft and Wizardry.

After data integration


HARRY POTTER
and the Philosopher's Stone
J.K.ROWLING
9¾
HOGWARTS EXPRESS
"...this is a terrific book." The Sunday Telegraph

- **Title**: Harry Potter and the Philosopher's Stone
- **Other title**: Harry Potter and the sorcerer's stone
- **Author**: J. K. Rowling
- **Illustrator**: Thomas Taylor
- **Publication year**: 1997
- **Publisher**: Bloomsbury (UK)
- **ISBN**: 0-7475-3269-9, 0747574472 (pbk), 9781338299144 (paperback), 133829914X (paperback)
- **Subjects:** Juvenile fiction, fantasy fiction, Wizards—Fiction, Magic-–Fiction, Schools-Fiction
- **Dewey**: 823.914
- **Summary**: Rescued from the outrageous neglect of his aunt and uncle, a young boy with a great destiny proves his worth while attending Hogwarts School for Witchcraft and Wizardry.

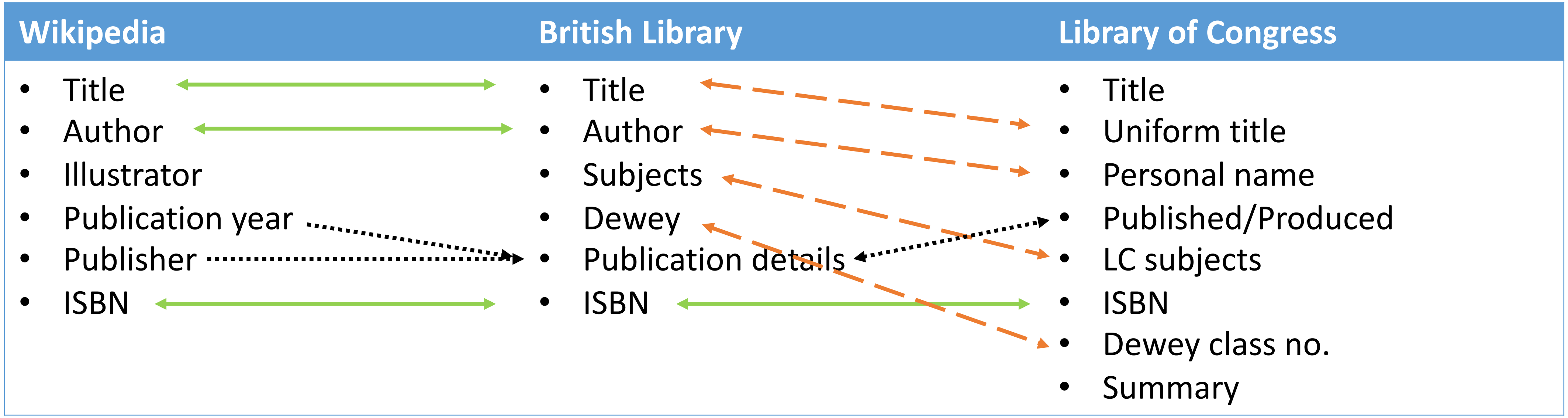
Data Integration

- Schema matching focuses on finding the correspondence among schema elements in two semantically correlated schemata
- Schema mapping describes how a source database schema relates to a target database schema
- Record linkage (also known as entity resolution and deduplication) identifies records that refer to the same logical entity
- Data fusion focuses on resolving conflicts and determining the true data values, leveraging information in heterogeneous data sources

X. Wang, L.M. Haas and A. Meliou, Explaining Data Integration, IEEE Data Eng. Bull. 41(2) (2018), 47–58

Data integration challenges

- Attributes may have different names-> entity attributes need to be **mapped**



Wikipedia	British Library	Library of Congress
• Title	• Title	• Title
• Author	• Author	• Uniform title
• Illustrator	• Subjects	• Personal name
• Publication year	• Dewey	• Published/Produced
• Publisher	• Publication details	• LC subjects
• ISBN	• ISBN	• ISBN
		• Dewey class no.
		• Summary

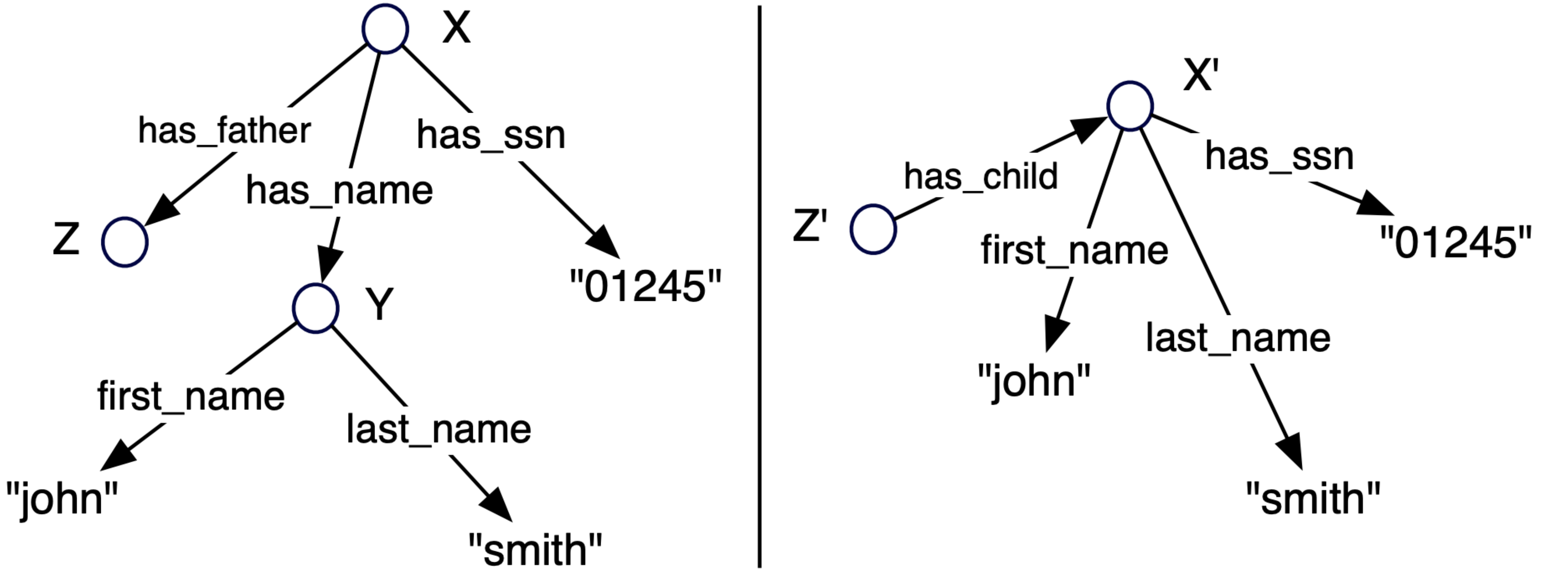
Data integration challenges

- Attribute values are not the same -> inconsistency needs to be resolved during **merging**

	Wikipedia	British Library	Library of Congress
Title	Harry Potter and the Philosopher’s Stone	Harry Potter and the Philosopher’s Stone	Harry Potter and the sorcerer’s stone
ISBN	0-7475-3269-9	0747574472 (pbk)	9781338299144 (paperback), 133829914X (paperback)
Subject		Juvenile fiction, fantasy fiction	Wizards—Fiction, Magic-–Fiction, Schools-Fiction
Dewey		823.914	823/.914 [Fic]
Publishing status	• Publication year: 1997 • Publisher: Bloomsbury (UK)	• Publication details: London : Bloomsbury, 1997 2004 printing.	• Published/Produced: New York, NY : Scholastic Inc., [2018]

Data integration challenges

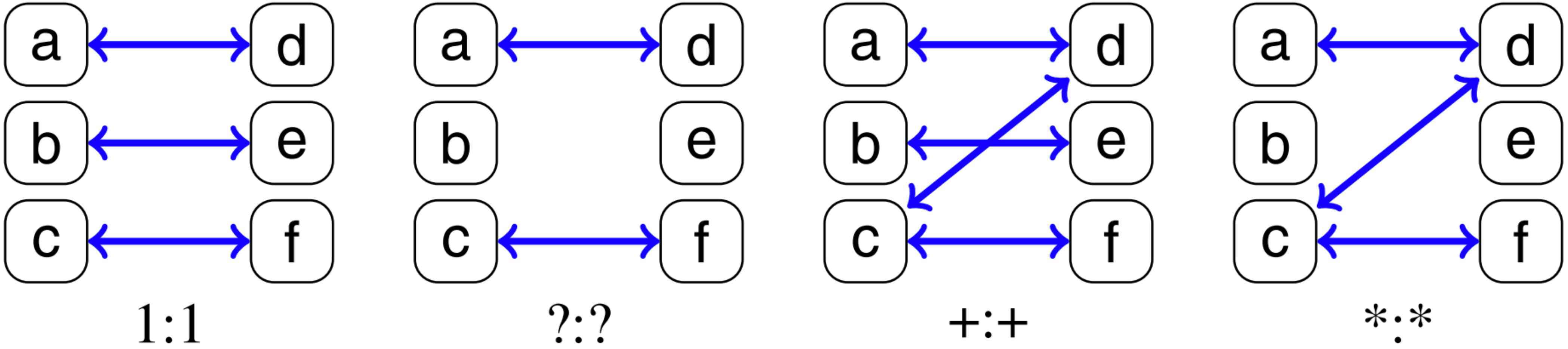
- Structural heterogeneity



Castano S., Ferrara A., Montanelli S., Varese G. (2011) Ontology and Instance Matching. https://doi.org/10.1007/978-3-642-20795-2_7

Data integration challenges

- Multiplicity
  - 1:1, 1:n, m:1, or m:n?
  - This applies to both entities and attributes



Euzenat, J. & Shvaiko, P. (2013). Ontology matching (2nd ed.). Springer Berlin Heidelberg.

Data integration challenges

- Matching score and threshold
  - <wiki:title> <bl:title> 1.0
  - <bl:Publication_details> <loc:Published_Produced> 0.7
  - False positive vs false negative

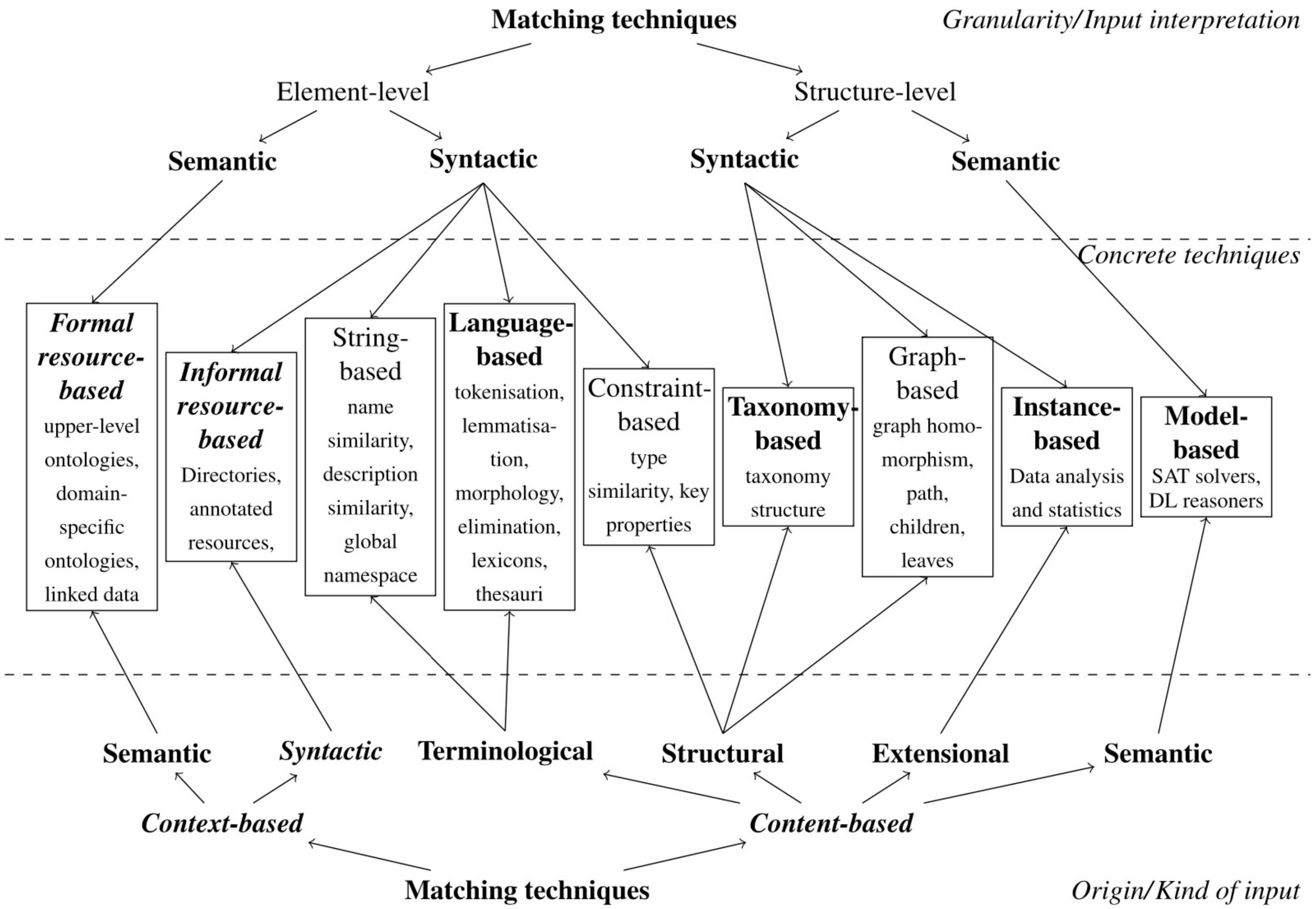
- Data sets are too big -> Scalability issues
  - Library of Congress has 19M+ book records, while British Library has 25M+ and WorldCat 400M+!
  - Optimisations
    - Reduction of the number of comparisons
    - Reduction of the cost of each comparison

Schema and instance matching

Ontology matching

- A schema is a structure of metadata describing how data, i.e., instances, can be stored, accessed, and interpreted by users and applications.
- Schema matching
  - Whether two concepts (e.g., *book* vs *manuscript*) or two properties (e.g. *birthplace* vs *hometown*) are the same
- Instance matching
  - Whether different instances refer to the same real-world entity in a given domain (e.g., a person, a place, a movie, a book, etc.)

Classification of schema matching techniques


Matching techniques
Granularity/Input interpretation
Element-level
Structure-level
Semantic
Syntactic
Syntactic
Semantic
Concrete techniques
Formal resource-based
upper-level ontologies, domain-specific ontologies, linked data
Informal resource-based
Directories, annotated resources,
String-based
name similarity, description similarity, global namespace
Language-based
tokenisation, lemmatisation, morphology, elimination, lexicons, thesauri
Constraint-based
type similarity, key properties
Taxonomy-based
taxonomy structure
Graph-based
graph homomorphism, path, children, leaves
Instance-based
Data analysis and statistics
Model-based
SAT solvers, DL reasoners
Semantic
Syntactic
Terminological
Structural
Extensional
Semantic
Context-based
Content-based
Matching techniques
Origin/Kind of input

Euzenat, J. & Shvaiko, P. (2013). Ontology matching (2nd ed.). Springer Berlin Heidelberg.

Categories of matching techniques

- Element-level techniques consider ontology entities or their instances in isolation from their relations with other entities or their instances
- Structure-level techniques consider the ontology entities or their instances to compare their relations with other entities or their instances.

Element-level techniques

- String-based: matching names or descriptions of entities
- Linguistic-based: use NLP, lexicons, or domain specific thesauri to match words based on linguistic relations (homonymy, synonymy, partonomy, etc) or exploiting morphological properties
- Constrained-based: take into account internal constraints applied to the definitions of entities, e.g. types, cardinality of properties, etc.
- Resource-based: use external upper-level or domain-specific ontologies to bridge classes to be matched.

Element-level techniques: string-based

- Levenshtein distance represents the number of insertions, deletions, and substitutions required to change one word to another.
  - For example, the Levenshtein distance between "kitten" and "sitting" is 3:
    - **k**itten → **s**itten (substitution of "s" for "k"),
    - sitt**e**n → sitt**i**n (substitution of "i" for "e"),
    - sittin → sittin**g** (insertion of "g" at the end).

- Damerau-Levenshtein distance counts transpositions as a single edit
  - For example, damerau_levenshtein_distance('fish', 'ifsh') == 1 while levenshtein_distance('fish', 'ifsh') == 2

Element-level techniques: string-based

- Hamming distance between two equal-length strings of symbols is the number of positions at which the corresponding symbols are different.
  - For example, the Hamming distance between "karoln" and "kerstn" is 3.

- N-gram
  - Takes as input two strings and calculate the number of the common n-grams between them, normalised by max(length(string1), length(string2))
  - Trigrams for matching is mat, atc, tch, chi, hin, ing

Element-level techniques: linguistic-based

- Tokenization
  - Parses names into tokens by recognizing punctuation, cases
  - string-based methods -> [string, based, methods]
- Lemmatisation
  - Analyses morphologically tokens to find their basic forms
  - methods -> method
- Remove stop words
  - a, he, them, by, from

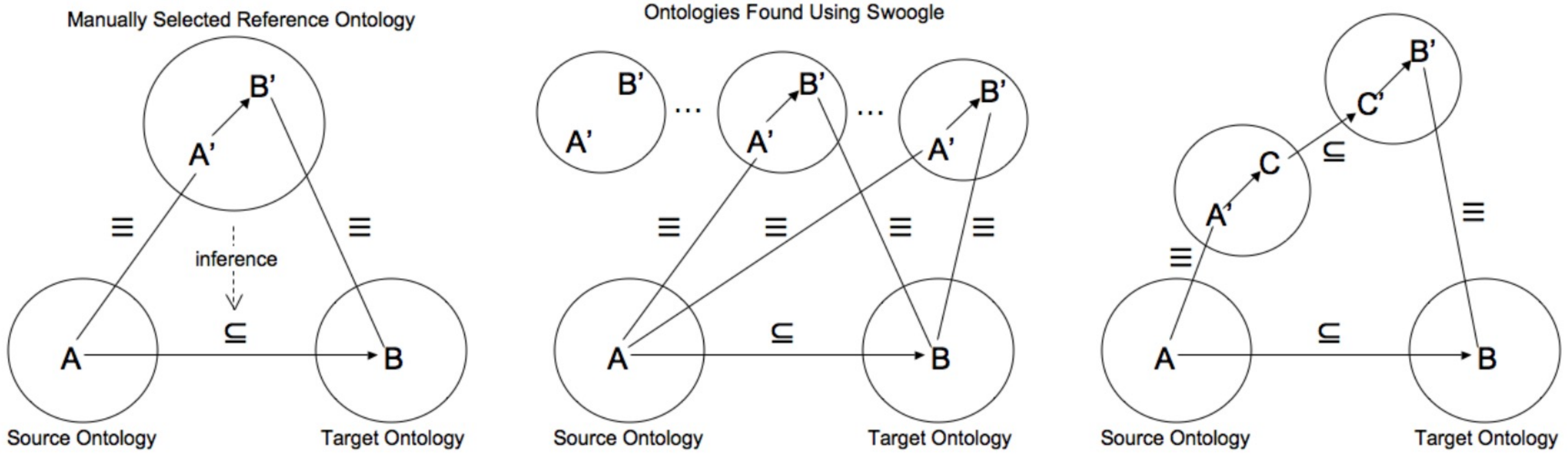
Element-level techniques: phonetic-based

- Phonetic encoding
  - The goal is for homophones to be encoded to the same representation so that they can be matched despite minor differences in spelling.
  - Soundex is a phonetic algorithm for indexing names by sound, as pronounced in English.
    - soundex(‘Ashcraft') == soundex(‘Ashcroft') == ‘A261’
    - Soundex('Rupert') == soundex('Robert') == 'R163’

Element-level techniques: resource-based

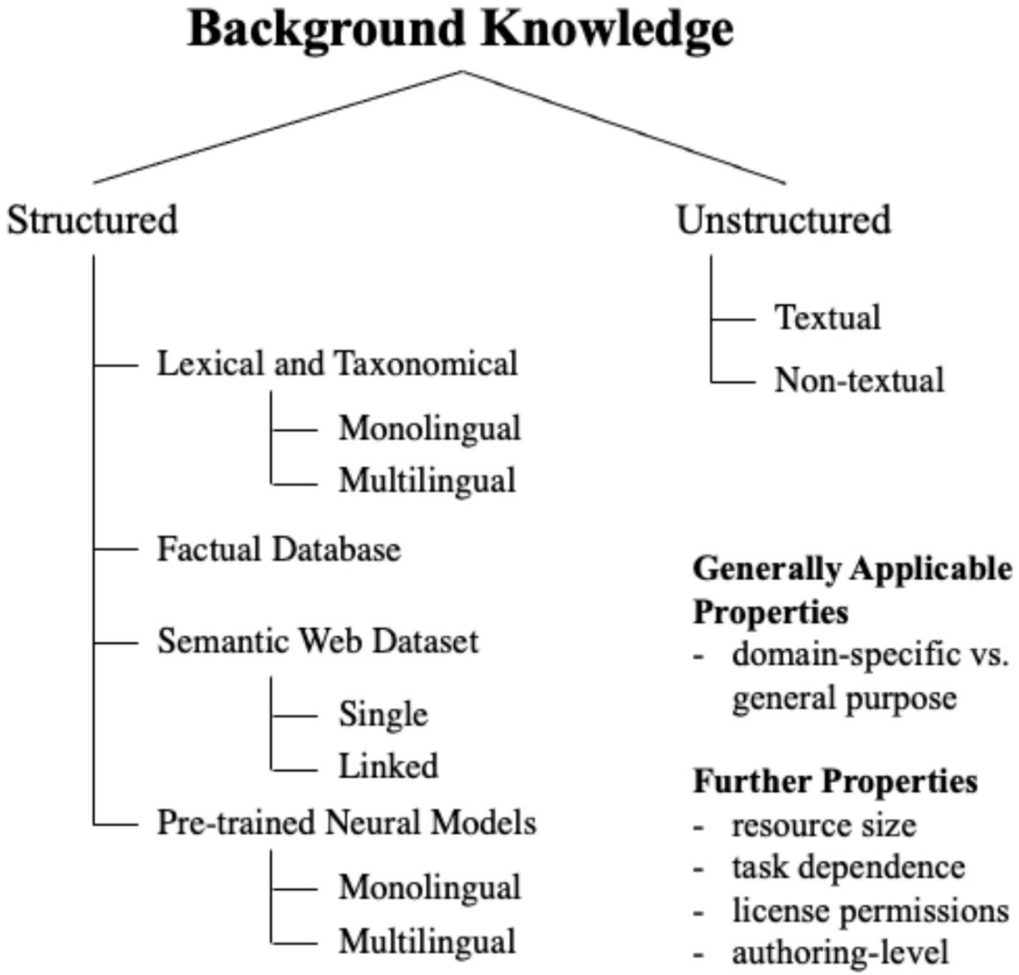
- Sense-based: WordNet
  - $A \sqsubseteq B$ if A is a hyponym or meronym of B
    - apple ⊑ fruit
  - $A \sqsupseteq B$ if A is a hypernym or holonym of B
    - animal ⊒ cat
  - $A = B$ if A and B are synonyms
    - student = pupil
  - $A \perp B$ if A and B are antonyms or the siblings in the part of hierarchy
    - giant ⊥ dwarf

Element-level techniques: resource-based



Sabou M., d'Aquin M., Motta E. (2008) Exploring the Semantic Web as Background Knowledge for Ontology Matching. In: Spaccapietra S. et al. (eds) Journal on Data Semantics XI. Lecture Notes in Computer Science, vol 5383. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-92148-6_6

Background knowledge-based matching



Most Used Background Knowledge Resources:

- WordNet
- Translation APIs
- Biomedical ontologies, e.g., UMLS, UBERON, BioPortal, MeSH

http://www.semantic-web-journal.net/content/background-knowledge-ontology-matching-survey-0

Structure-level techniques

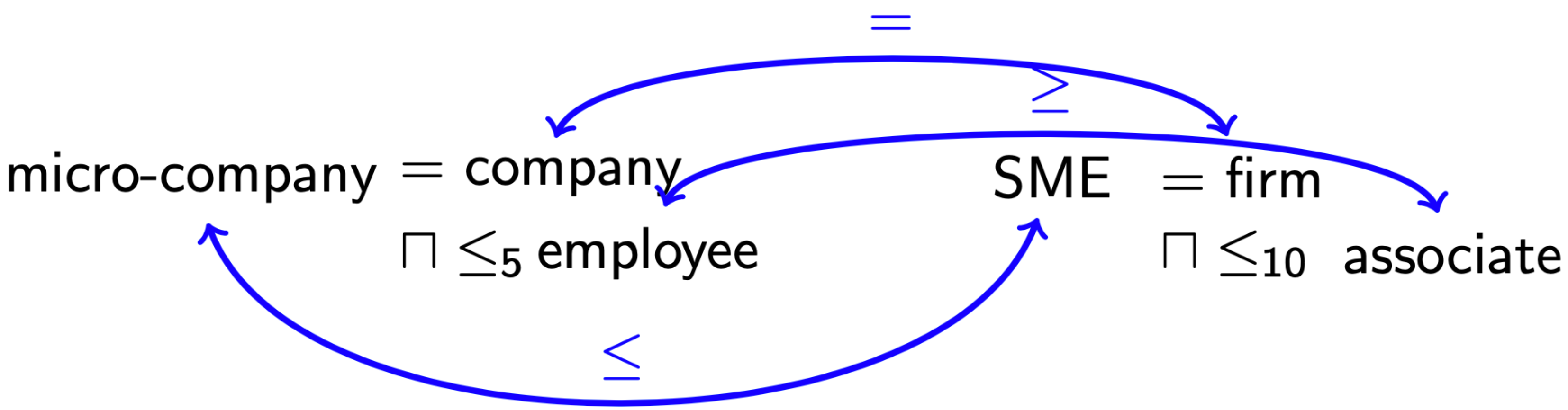
- Graph-based: consider ontologies as labelled graphs, assumption: if nodes are similar then their neighbours must be similar
- Taxonomy-based: consider only specialisation/generalization relation
- Model-based: take into account semantic interpretation of the ontologies, assumption: if two entities are the same then they share the same interpretation
- Instance-based: use individual representation of classes, i.e., classes are considered similar if they share many instances

Structure-level techniques: graph-based

- Children
  - Two non-leaf schema elements are structurally similar if their immediate children sets are highly similar
- Leaves
  - Two non-leaf schema elements are structurally similar if their leaf sets are highly similar, even if their immediate children are not
- Iterative fix point computation
  - If the neighbours of two nodes of the two ontologies are similar, they will be more similar

Structure-level techniques: model-based

Description Logics (DL)-based

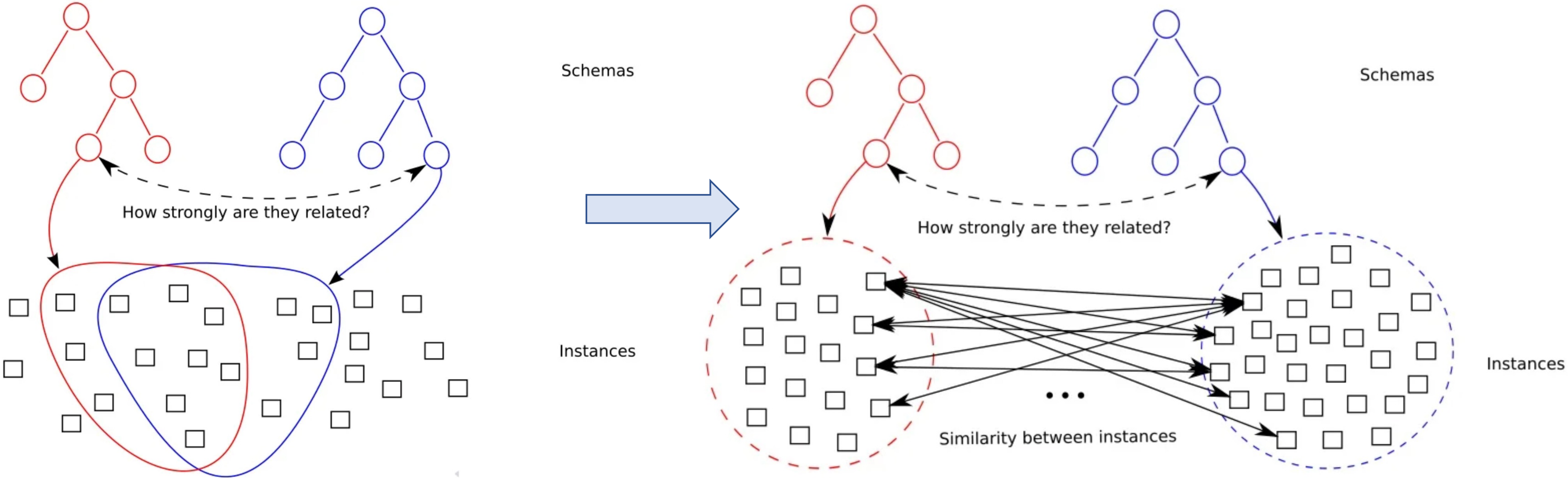


$\text{micro-company} = \text{company} \sqcap \leq_5 \text{employee}$

$\text{SME} = \text{firm} \sqcap \leq_{10} \text{associate}$

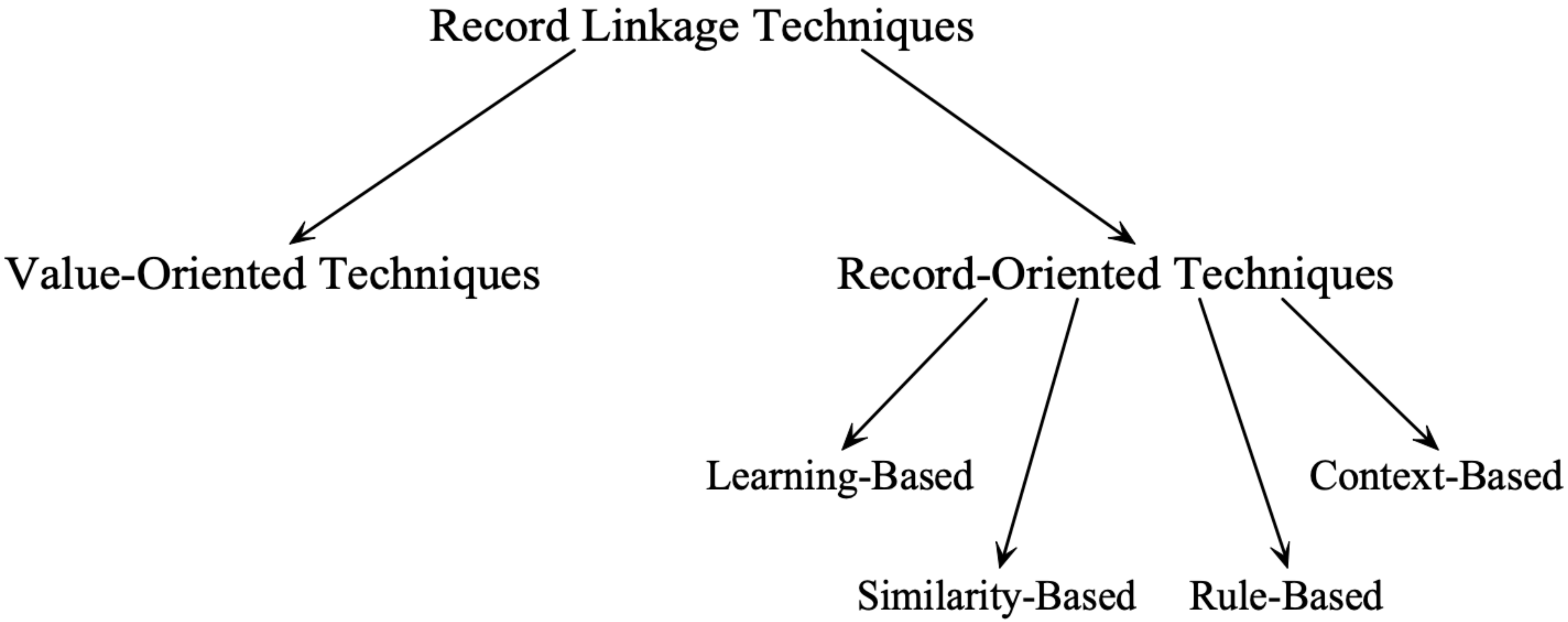
$$\frac{\text{company} = \text{firm} \; ; \; \text{associate} \sqsubseteq \text{employee}}{\text{micro-company} \sqsubseteq \text{SME}}$$

Structure-level techniques: instance-based


Schemas
How strongly are they related?
Instances
Schemas
How strongly are they related?
• • •
Similarity between instances
Instances

Wang S., Englebienne G., Schlobach S. (2008) Learning Concept Mappings from Instance Similarity. In: Sheth A. et al. (eds) The Semantic Web - ISWC 2008. ISWC 2008. Lecture Notes in Computer Science, vol 5318. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-88564-1_22

Instance matching (record linkage)


Record Linkage Techniques
Value-Oriented Techniques
Record-Oriented Techniques
Learning-Based
Context-Based
Similarity-Based
Rule-Based

Castano S., Ferrara A., Montanelli S., Varese G. (2011) Ontology and Instance Matching. https://doi.org/10.1007/978-3-642-20795-2_7

Value-oriented techniques

- Assumption: the similarity level of two records (entities) can be derived by matching the values of their comparable attributes

- Mostly focus on similarity of string attributes
  - Character-based: Edit Distance, Smith-Waterman Distance, Jaro Distance
    - Typographical variations, e.g. “organisation” vs “organization”
  - Token-based: Cosine similarity over TF-IDF vectors or word embeddings
    - Different conventions for describing data, e.g. “J.K. Rowling” vs “Rowling, J. K.”
  - Phonetic-based: Soundex, NYSIIS, Metaphone
    - Phonetic similarity, e.g. “Kageonne” vs “Cajun”

Classification of matching vs non-matching

- When the similarity of each pair of corresponding attribute values is computed, a decision engine is needed to classify whether two entities match or not
    - Learning-based
    - Similarity-based
    - Rule-based
    - Context-based

Learning-based techniques

- Supervised learning
  - Training data: a set of instance pairs and their expected classification (i.e. matching or non-matching records)
  - Non-trivial to get a high-quality and balanced training data set
  - Manually adding ambiguous cases if possible

- Unsupervised learning
  - Clustering record pairs with similar features that belong to the same class (i.e. matching or non-matching records)

- Combining different learning techniques

Similarity-based techniques

- Considering the input records as long attribute values
  - Concatenate all attribute values into one single string

- Average similarity of each pair of corresponding attribute values
  - Weighted average may better reflect domain knowledge

- Taking into account the frequency each value occurs
  - A matched "Paul Smith" counts less than the matched "Bamidele Melisizwe"

Rule-based techniques

- Instead of similarity values, a Boolean output is assigned
  - For example, if two records denoting books share the same value on attributes “Title”, “Author” and “Publication year”, there is a very high probability that the considered records refer to the same book.

Context-based techniques

- Taking into account the relationships between records
  - For example, matching clusters of records instead of individual records

Optimisation

- Reduction of the number of comparisons
  - Blocking: dividing instances in homogeneous and mutually exclusive *blocks* and comparing only with instances belonging to the same block
  - Sorted neighbourhood: sorting instances according to the value they assume on the property with the highest discriminating power and only comparing instances within a shifting- window of a fixed dimension

- Reduction of the cost of each comparison
  - Only a subset of the corresponding attribute values matters
    - E.g. title-author-year is a reasonable subset to distinguish books

Similarity filter and alignment extraction

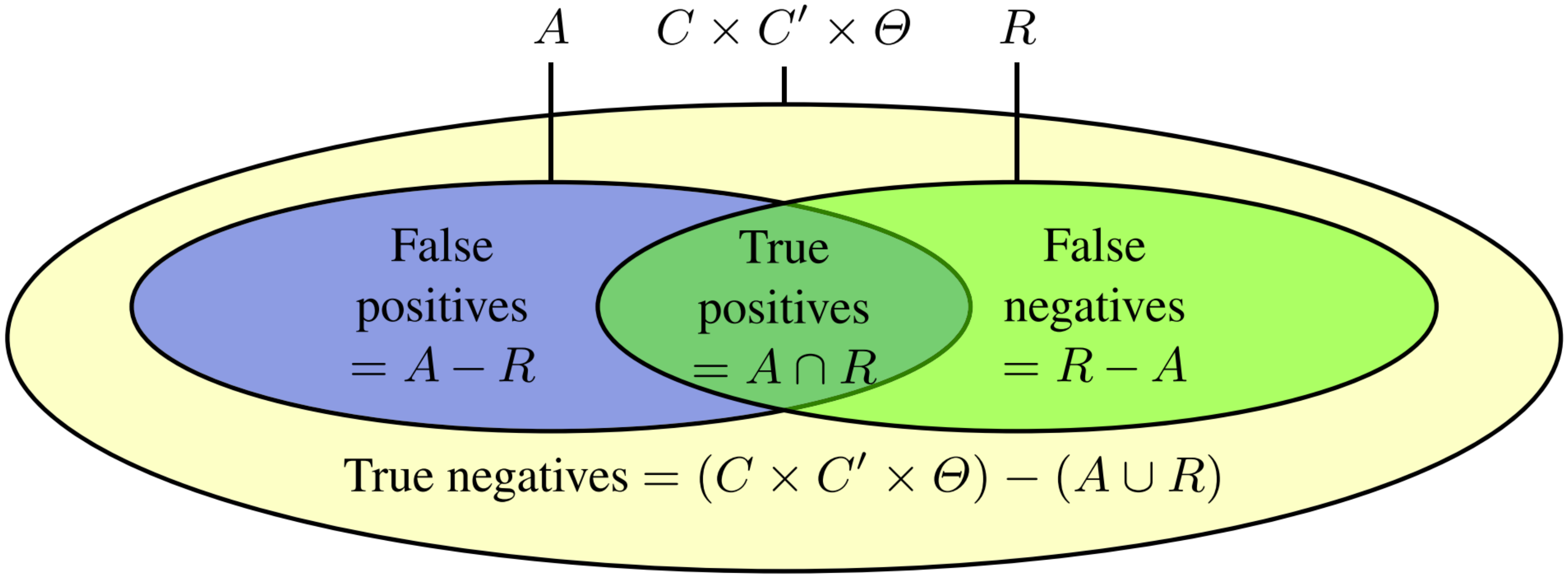
- Many algorithms are based on similarity or distance computation. A few operations can be based on similarity/distance matrices.
- Various thresholding options to filter similarities
  - Hard threshold retains all the correspondence above threshold n;
  - Delta threshold consists of using as a threshold the highest similarity value out of which a particular constant value d is subtracted;
  - Gap threshold retains the correspondences ordered by decreasing similarity until the difference in similarity between two correspondences becomes larger than n;
  - Proportional threshold consists of using as a threshold the percentage of the highest similarity value;
  - Percentage retains the n% correspondences above the others.

Evaluation

Alignment quality measures

- Quality measures are the intrinsic measures of the alignment (they do not depend on any reference):
  - threshold on confidence or average confidence,
  - cohesion measures between matched entities, i.e., their neighbours are matched with each other,
  - ambiguity degree, i.e., proportion of classes matched to several other classes,
  - agreement or non-disagreement between the aligned ontologies,
  - violation of some constraints, e.g., acyclicity in the correspondence paths,
  - satisfaction of syntactic anti-patterns,
  - consistency and coherence.

Evaluation with a reference alignment


A
$C \times C' \times \Theta$
R
False positives $= A - R$
True positives $= A \cap R$
False negatives $= R - A$
True negatives $= (C \times C' \times \Theta) - (A \cup R)$

- Precision measures the ratio of correctly found matches (true positives) over the total number of returned matches (true positives and false positives)
- Recall measures the ratio of correctly found matches (true positives) over the total number of expected matches (true positives and false negatives)
- F-measure = 2 * P * R / (P + R)

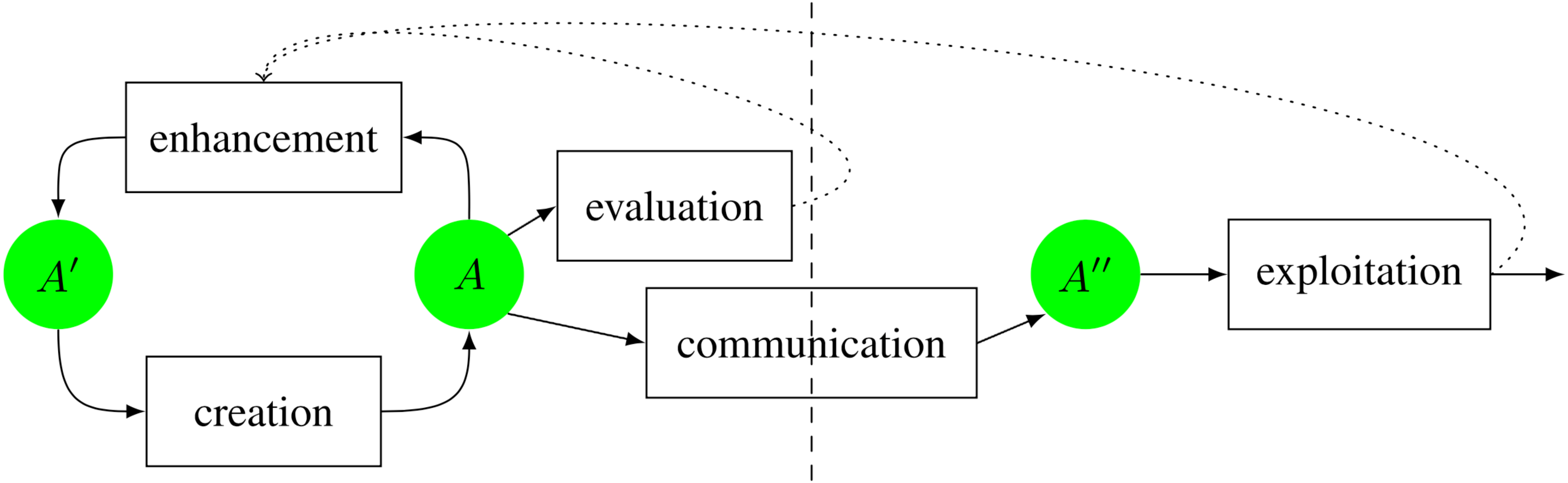
Euzenat Jérôme, & Shvaiko, P. (2013). Ontology matching (2nd ed.). Springer Berlin Heidelberg.

Evaluation without a reference alignment

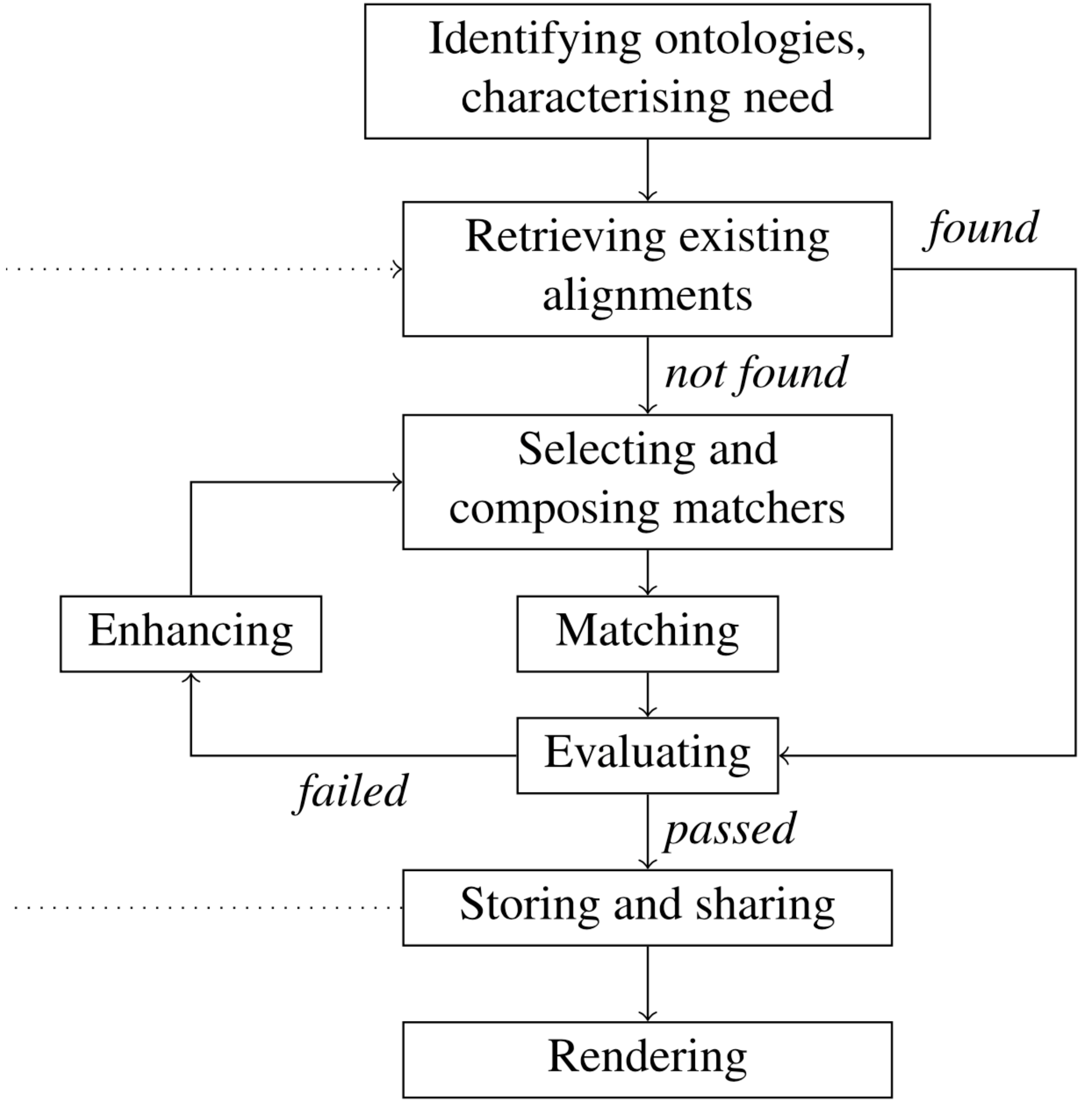
- Sampling or pooling-based evaluation
  - Build a subset of alignment using sampling or pooling
  - Expert evaluation
- End-to-end evaluation
  - Testing the quality of alignment using an application
- Performance measures
  - Speed, memory, scalability

Alignment lifecycle

The alignment life cycle



The matching methodology workflow


Identifying ontologies, characterising need
Retrieving existing alignments
found
not found
Selecting and composing matchers
Enhancing
Matching
Evaluating
failed
passed
Storing and sharing
Rendering

Learning and tuning

- A multi-strategy learning approach is useful when several learners are used, each one handling a particular kind of pattern that it learns best.
- Tuning refers to the process of adjusting a matcher for a better functioning in terms of:
  - better quality of matching results, measured, e.g., through precision or F-measure, and
  - better performance of a matcher, measured through resource consumption, e.g., execution time, main memory.

Summary

- Heterogeneity of ontologies is in the nature of the semantic web;
- Ontology matching is part of the solution;
- It can be based on many different techniques;
- Iterative process towards improved performance;
- There remain serious challenges ahead.

Challenges

- Large-scale and efficient matching,
- Matching with background knowledge,
- User involvement,
- Multilinguality,
- Social and collaborative matching,
- Uncertainty in matching,
- Reasoning with alignments,
- Alignment management.
- ...