Mitigating linked data quality issues in knowledge-intense information extraction methods

Albert Weichselbraun
Swiss Institute for Information Research
University of Applied Sciences Chur
Chur, Switzerland 7000
albert.weichselbraun@htwchur.ch

Philipp Kuntschik
Swiss Institute for Information Research
University of Applied Sciences Chur
Chur, Switzerland 7000
philipp.kuntschik@htwchur.ch

ABSTRACT

Advances in research areas such as named entity linking and sentiment analysis have triggered the emergence of knowledge-intensive information extraction methods that combine classical information extraction with background knowledge from the Web. Despite data quality concerns, linked data sources such as DBpedia, GeoNames and Wikidata which encode facts in a standardized structured format are particularly attractive for such applications.

This paper addresses the problem of data quality by introducing a framework that elaborates on linked data quality issues relevant to different stages of the background knowledge acquisition process, their impact on information extraction performance and applicable mitigation strategies. Applying this framework to named entity linking and data enrichment demonstrates the potential of the introduced mitigation strategies to lessen the impact of different kinds of data quality problems. An industrial use case that aims at the automatic generation of image metadata from image descriptions illustrates the successful deployment of knowledge-intensive information extraction in real-world applications and constraints introduced by data quality concerns.

CCS CONCEPTS

• **Information systems** → Incomplete data; Inconsistent data; Extraction, transformation and loading; Data cleaning; Entity resolution; • **Computing methodologies** → Information extraction;

KEYWORDS

linked data quality, mitigation strategies, information extraction, named entity linking, semantic technologies, applications

ACM Reference format:
Albert Weichselbraun and Philipp Kuntschik. 2017. Mitigating linked data quality issues in knowledge-intense information extraction methods. In *Proceedings of WIMS '17, Amantea, Italy, June 19-22, 2017,* 12 pages.
DOI: 10.1145/3102254.3102272

1 INTRODUCTION

Information Extraction (IE) is concerned with the automatic extraction of structured knowledge from unstructured or semi-structured text documents. Historically, IE methods focused on knowledge-poor approaches that apply machine learning, statistical analysis and natural language processing to the input text without leveraging knowledge from external sources.

Recent developments in research areas such as named entity linking and sentiment analysis have triggered the emergence of a new class of knowledge-rich IE methods that aspire toward combing classical IE techniques with background knowledge from the Web. These approaches have been proven to be very effective, especially for complex information extraction tasks such as (i) named entity linking that identifies and links mentions of named entities to a knowledge base, (ii) target sentiment analysis which assigns sentiment values to entities towards which a sentiment is expressed (such as products, persons and organizations), and (iii) aspect-based sentiment analysis that aims at identifying the reasons (e.g. battery life, display quality and usability of the software for a smartphone) for positive and negative sentiment assessments.

Linked data plays a pivotal role in this development, since it provides access to billions of triples that encode background knowledge relevant to a multitude of domains and tasks. High levels of standardization in terms of query protocols, serialization formats and data models foster interoperability and allow reuse of knowledge mining components that provide background knowledge to information extraction methods.

1.1 Motivation

Although there are many theoretical benefits of exploiting linked data for information extraction, realizing these advantages is still a challenging task since it requires (i) locating and mining relevant linked data sources, and (ii) adapting information extraction methods to exploit the mined knowledge.

Both steps are considerably complicated by the sheer size and diversity of the linked data ecosystem and heterogeneous data quality standards. Nevertheless, these problems are already well known from the World Wide Web and call for strategies towards addressing them.

This work, therefore, focuses on linked data quality issues relevant to information extraction and on strategies for addressing them. After introducing quality dimensions and mitigation strategies, we discuss these strategies based on an industry project that focuses on the automatic annotation of images drawing upon background knowledge from publicly available sources such as DBpedia, Wikidata and GeoNames.

WIMS '17, Amantea, Italy

978-1-4503-5225-3/17/06...$15.00
DOI: 10.1145/3102254.3102272

1.2 Contributions

Related work primarily focuses on taxonomies for linked data quality [35] as well as on automatic means for assessing quantitative aspects of linked data quality [8, 23, 27, 34]. The research presented in this paper extends and complements this work by (i) discussing the impact of data quality issues on using linked data to integrate background knowledge in information extraction methods (Section 3), (ii) investigating mitigation strategies for the corresponding linked data quality dimensions (Section 3), (iii) introducing information extraction, data enrichment and semantic search methods that apply these strategies on linked data obtained from DBpedia, Wikidata and GeoNames (Section 4), and (iv) presenting a real-world use cases that demonstrates the use of linked data for information extraction and elaborates design choices - based on the discussed mitigation strategies - that help in lessen the impact of linked data quality issues (Section 5).

2 STATE OF THE ART

The following discussion of related work first elaborates on the role of information extraction in Web Intelligence and Big Data applications and then shortly discusses how information extraction benefits from background knowledge. Finally, we provide an overview of research on linked data quality issues.

2.1 Information Extraction for Web Intelligence and Big Data Applications

Information extraction plays a key role in many Web Intelligence and Big Data applications, since it provides powerful methods for analyzing textual content in online and social media. The potential and capabilities of these analytics have been successfully demonstrated in many domains such as politics [21], environmental communication [24], financial market analysis [5], health care [33] and marketing [32].

Chung and Zeng [6], for instance, use network and sentiment analysis on Twitter to investigate the discussion on the U.S. immigration and border security. The authors use graph mining to uncover major phases in the Twitter coverage, identify opinion leaders and influential users, and draw upon information extraction to investigate the differences in sentiment, emotion and network characteristics between these phases. Scharl et al. [24] present visual tools and analytics to support environmental communication in the Media Watch on Climate Change (www.ecoresearch.net/climate), the Climate Resilience Toolkit (toolkit.climate.gov) and the NOAA Media Watch [25]. All three platforms aggregate and analyze the coverage of environmental topics in different outlets, including news media, Fortune 1000 companies and social media such as Twitter, Facebook, Google+ and YouTube. Li et al. [15] draw upon company-specific news articles to study their impact on the movements of stock markets. They conclude that public sentiments voiced in these articles cause fluctuations of the market, although their impact depends on the company as well as the article content. Kim et al. [13] investigate the coverage of the Ebola Virus on Twitter and in news media. They apply topic modelling and named entity extraction to create topic and entity networks, compute per-topic sentiment scores and analyze the temporal evolution of these networks.

2.2 How Information Extraction Benefits from Background Knowledge

Named entity linking by its very nature relies on knowledge sources to which mentions of named entities are linked.

Most named entity linking approaches either directly use document collections such as Wikipedia or draw upon linked open data sources such as DBpedia, YAGO and GeoNames as knowledge base.

Wikipedia, for instance, has been used for training named entity linking models [12, 18] as well as for improving the accuracy of the disambiguation process [10, 11, 20]. Work by Han and Zhao [10] suggests that integrating semantic knowledge from Wikipedia improves disambiguation by as much as 10.7% over traditional bag-of-word approaches, and by 16.7% over traditional social network-based disambiguation methods.

The second group of approaches directly draws upon linked open data repositories. DBpedia Spotlight [7] is one of the best known, publicly available tools for extracting named entities from textual content. AIDA [11] leverages data from multiple linked open data sources and uses links between entities to compute the context similarity and coherence for disambiguation. AIDA light builds upon this research but focuses on reducing the memory and resource footprint of the named entity linking component[17]. Recognyze [31] demonstrates the flexibility of information extraction approaches that build upon linked data technology. The component features comprehensive graph mining and preprocessing capabilities and is able to use arbitrary linked data repositories as knowledge sources. The advantages of this design choice became apparent when the component which has originally been developed within an industry project and optimized for proprietary linked enterprise data on companies and persons has been adapted to linked open data sources such as DBpedia, GeoNames and YAGO within hours.

Natural language processing (NLP) is another field that could benefit tremendously from integrating background knowledge. NLP researchers are particularly interested in common sense knowledge (which humans acquire during the formative years of their lives) and common knowledge (which people continue to gather in their everyday lifes) [4] since they help in better understanding the context of words and statements expressed in natural language, paving the way from bag-of-words approaches to more complex bag-of-concepts approaches that overcome the problem of ambiguity. Cambria and White [4] even predict that future research will replace bag-of-concepts with more sophisticated bag-of-narratives that represent interconnected stories and episodes within the text.

Knowledge resources such as ConceptNet (conceptnet5.media-.mit.edu) [26], SenticNet (sentic.net) [3] and SentiWordNet (sentiwordnet.isti.cnr.it) [1] enable this evolution of natural language processing techniques. Recent research in sentiment analysis, for example, demonstrates how background knowledge can be used to distinguish between ambivalent concepts [29], and consider common and common sense knowledge in the extraction of sentiment targets and aspects [28] .

The benefits of knowledge obtained from linked data are not limited to the information extracting process but rather relevant for the whole information extraction life cycle. For instance, linked data enables statistics on the extracted named entities, and visualizations of their interactions with each other in so called entity maps [25].

2.3 Linked Data Quality Issues

DBpedia publishes structured information from Wikipedia that is extracted by automatic exploration tools together with manually crafted property mappings. As human labor takes a major part in both, creating content in the underlying semi-structured knowledge base as well as the mapping into machine readable properties [14], and human performance by its very nature is individual, there are many levels where quality issues can arise.

Mihindukulasooriya et al. [16] show that especially local versions of DBpedia lack in conciseness, consistency, syntactic validity and semantic accuracy of multiple properties. They attribute this problem to the extraction process and the use of inadequate automatic property generation methods instead of hand-crafted template mappings. Inconsistent capitalization, use of accents, spelling mistakes, wrong domain values, and properties that are simultaneously used as object and datatype properties lead to quality issues for both, property description and -content. Thakkar et al. [27] assess and compare data quality in DBpedia and Wikidata in regard to question answering use cases. They conclude that Wikidata offers a higher data quality in terms of relevant metrics for this particular use cases, although their study only focused on statistically coverage and did not check on validity or defectiveness of the data itself.

The need to evaluate linked data quality has led to the emergence of (semi-)automatic quality assessment tools and frameworks. Luzzu [8], for example, provides a scalable, extensible, interoperable, and customisable framework that allows setting individual weights for each metric, and therefore ranks and compares datasets based on the requirements for the intended use case. Ruckhaus et al. [23] introduce LiQuate which combines Bayesian Networks and rule-based systems to identify ambiguities, suggest possible inconsistencies, and identify potential incompleteness in linked data. TripleCheckMate, in contrast, integrates a manual crowd-sourcing oriented process which is used to assemble a gold standard of dataset errors, with a semi-automatic process that trains a machine learning component with the gold standard. The created classifier then identifies further triples with a high probability of being incorrect [34].

Zaveri et al. [34] conduct a systematic review of existing approaches to assess data quality of linked open data, analyzing 21 relevant papers that have been published between 2002 and 2012. Their aim is to create a clear understanding of linked data quality in terms of quality dimensions, metrics, type of data and tools available [35].

Data quality issues can be addressed on multiple levels. Paulheim and Bizer [19] propose SDType and SDValidate as two methods that rely on the statistical distribution of types and properties in a linked data repository. These methods add missing type information and identify potentially incorrect statements that have been generated during information extraction tasks which created the linked data repository. Zaveri et al. [34] classify the data quality issues based on error correction strategies into errors that can be solved by (i) amending the extraction framework, (ii) correcting the property mapping, or (iii) adjusting the semi-structured knowledge base, Ristoski and Paulheim [22] discuss how linked data can be applied to data mining and knowledge discovery. They also suggest to deal with data problems in a separate preprocessing step that handles missing values, identifies incorrect data, eliminates duplicates and performs conflict resolution.

This paper complements related research that investigates methods for describing, quantifying or improving the quality of linked data sources, by focusing on strategies for mitigating data quality issues in linked data to facilitate their use as background knowledge source for knowledge-intensive information extraction processes.

3 LINKED DATA QUALITY DIMENSIONS AND MITIGATION STRATEGIES

Linked data offers standardized protocols such as SPARQL to enable easy access to data snippets. The consistent use of namespaces, unique identifiers and links between data sources enables interoperability between different datasets, promotes sharing of vocabulary and, therefore, more compatible data models. Although sources such as DBpedia are based on unstructured data, extractors for these sources and the created artifacts are shared within the community, attracting error reports and fixes that improve data quality, and stimulate the reuse of extraction components.

This section draws upon the conceptual data quality framework introduced by Zaveri et al. [35] for qualifying these data quality issues. The framework distinguishes 18 different quality dimensions that are grouped in

- *Accessibility dimensions* covering quality problems related to accessing the data such as availability, licensing and interlinking;
- *Intrinsic dimensions* that are independent of the application and evaluate whether the provided information is correct, consistent, concise and complete;
- *Contextual dimensions* capturing the data's fitness for use in a particular application or use case; and
- *Representational dimensions* that assess the quality of the chosen data design such as interoperability and interpretability.

The usability of linked data for information extraction depends considerably on how the data has been created, dataset specific restrictions and the proficiency of the dataset's publisher. DBpedia, for instance, is created by extracting structured knowledge from Wikipedia and, therefore, [14] inherits the strengths and shortcomings of the underlying data source, in addition to errors that have been introduced during the knowledge extraction process. Nevertheless, it is one of the most comprehensive datasets and, therefore, highly popular.

It is also important to note that data quality may vary considerably between sources and even between different releases of the same datasource. The quality of DBpedia, for instance, often changes due to modifications in the extractors but also due to the addition or removal of datasets. Therefore, testing the data quality of new dataset releases is highly recommended.

Figure 1 illustrates the data quality dimensions that are of a particular importance for the selection and preprocessing steps as outlined by Fayyad et al. [9]. *Selection* refers to the process of choosing datasets that are relevant to the particular application, and loading these datasets into a task-specific local linked data repository. *Preprocessing* comprises all steps necessary to use these

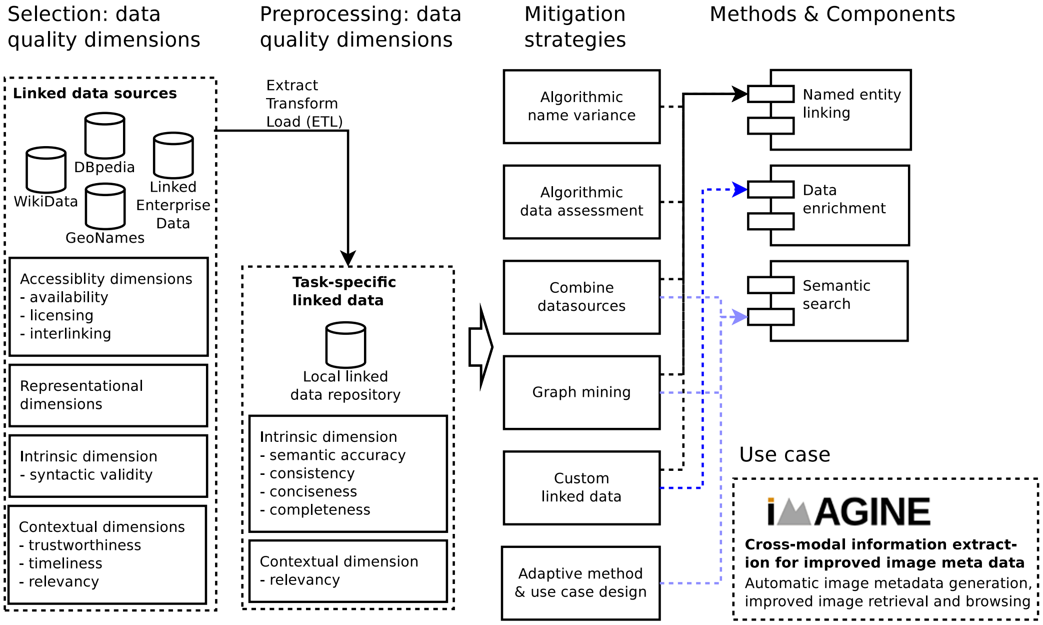


Figure 1: Linked data quality dimensions and mitigation strategies.

data in the subsequent information extraction tasks but also by components that draw upon the extracted data such as data enrichment and semantic search. The chosen use case hereby determines the methods used, useful linked data sources and relevant data quality dimensions.

For instance, since the framework focuses on industrial use cases task-specific local linked data repositories are used for performance and reliability reasons. Therefore data quality dimensions such as security and performance which apply to querying third party SPARQL endpoints are no longer relevant in this setting.

3.1 Data Quality in the Selection Process

In our experience, the potential impact of data quality problems needs to be addressed already at the *selection* step where potential linked data sources that could support information extraction methods are chosen. Domain experts, therefore, need to

(1) decide on linked data sources relevant to the use case, information extraction components, and methods used to leverage the extracted information (semantic search and visual analytics in the example), and
(2) investigate their data quality, its impact on the use case and mitigation strategies for addressing them.

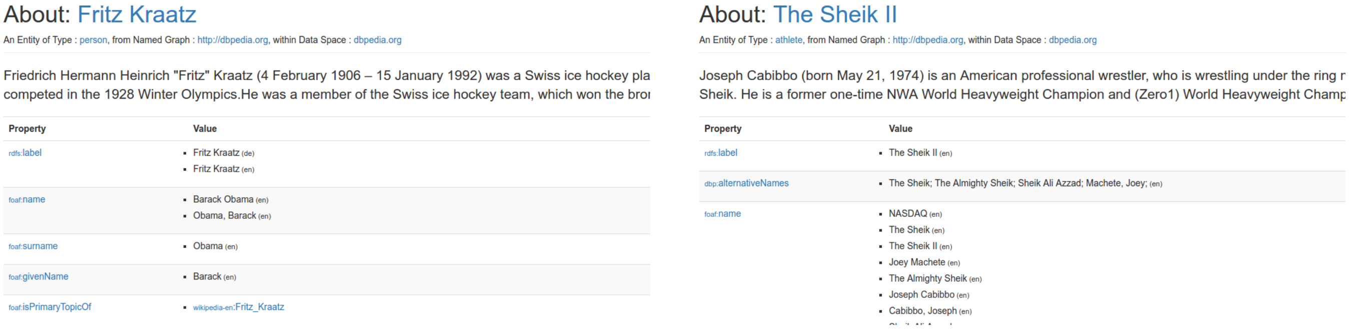
In this step accessibility dimensions and contextual dimensions determine the extend to which data is usable and relevant for the chosen task. For example, datasets with severe licensing restrictions, that include untrustworthy data sources or do not contain knowledge relevant to the given use case will be excluded in this early stage.

An Extract, Transform and Load (ETL) process mitigates many representational and contextual data quality issues as well as some intrinsic data quality problems by loading chosen subgraphs from public datasets into a local linked data repository that will be used as background knowledge for information extraction. Extract refers to the selection of subgraphs relevant to the application, transform to cleanup steps required for addressing data quality issues, and load to uploading the data to a local repository. The representational data quality dimensions and the intrinsic dimension syntactic validity determine the effort required for these data preparation processes.

Domain experts need to decide upfront on relevant datasets and whether the data quality in terms of *timeliness*, i.e. the recency of the available data, is sufficient. DBpedia, for example, is created by extracting structure knowledge from Wikipedia and currently features a half-yearly release cycle which causes a considerable time-lag between changes to Wikipedia and their propagation to DBpedia statements. DBpedia's online version, for instance, still listed Donald Trump as presidential candidate in March 2017.

Although many linked open data projects provide data dumps in standard formats such as Turtle, N3 or XML/RDF, data problems are quite common when importing these data. Even data dumps provided by popular lighthouse projects often require a prior transformation step as outlined below:

(1) Representational data quality: GeoNames, for example, uses a non-standard format, where every resource is serialized with its URL, followed by the corresponding XML/RDF snippet in the next line. Importing this format required transformation of the data snippets to representations that can be imported directly into SPARQL repositories. DBpedia, in contrast, provides much more user-friendly downloads that are available in multiple standard-compliant formats, separated by language and dataset (article texts, categories, links to other datasets, etc.)



About: Fritz Kraatz

An Entity of Type : person, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

Friedrich Hermann Heinrich "Fritz" Kraatz (4 February 1906 – 15 January 1992) was a Swiss ice hockey pla
competed in the 1928 Winter Olympics.He was a member of the Swiss ice hockey team, which won the bro

Property	Value
rdfs:label	▪ Fritz Kraatz (de) ▪ Fritz Kraatz (en)
foaf:name	▪ Barack Obama (en) ▪ Obama, Barack (en)
foaf:surname	▪ Obama (en)
foaf:givenName	▪ Barack (en)
foaf:isPrimaryTopicOf	▪ wikipedia-en:Fritz_Kraatz

About: The Sheik II

An Entity of Type : athlete, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

Joseph Cabibbo (born May 21, 1974) is an American professional wrestler, who is wrestling under the ring r
Sheik. He is a former one-time NWA World Heavyweight Champion and (Zero1) World Heavyweight Champ

Property	Value
rdfs:label	▪ The Sheik II (en)
dbp:alternativeNames	▪ The Sheik; The Almighty Sheik; Sheik Ali Azzad; Machete, Joey; (en)
foaf:name	▪ NASDAQ (en) ▪ The Sheik (en) ▪ The Sheik II (en) ▪ Joey Machete (en) ▪ The Almighty Sheik (en) ▪ Joseph Cabibbo (en) ▪ Cabibbo, Joseph (en)

Figure 2: DBpedia error caused by a failure in the knowledge extraction (left) versus a suspicious entry that is correct (right). Both entries have been retrieved on 14 March 2017.

(2) An example for an *interoperability data quality* issue becomes apparent when using Wikidata PageRank information (people.aifb.kit.edu/ath) with Wikidata. Since the page rank dataset uses the `dbpediawikidata` rather than the `wikidata` namespace which is only linked to `dbpedia-wikidata` by a `owl:sameAs` property, a conversion between these namespaces is required, unless the application supports reasoning.

(3) *Interlinking* is an enabler for interoperability and refers to the degree to which resources representing the same concepts are linked to each other. WordNet 3.0 RDF (semanticweb.cs.vu.nl/lod/wn30/), for instance, does not provide links to other linked data vocabularies and rarely reuses properties although a mapping for some WordNet-specific properties to SKOS exists. Missing links between datasets seriously restrict the ability to combine datasets unless these links are created in a preprocessing step.

(4) *Syntactic validity* is an intrinsic data quality dimension. Wikidata provides dumps in N-Triple format, but exports resource URLs with special characters such as double quotes, carets, curled brackets, tabulators etc. in a format that is considered syntactically incorrect by popular RDF repositories such as Apache Jena and RDF4J. Users, therefore, need to use transformation scripts that correctly escape and clean these resource URLs before they can load these data into a repository. A number of GitHub projects are dedicated to fixing such issues, although the sustainable solution would be ensuring syntactic validity in the created data dumps.

In practice transformation steps, as the ones described above, are often necessary before linked data can be loaded into repositories, although compatible data representations and standardized export formats could considerably cut down the transformation effort.

3.2 Data Quality in the Preprocessing Step

Preprocessing queries background knowledge from the local linked data repository and performs further mitigation steps to address intrinsic data quality issues such as semantic accuracy, consistency and contextual dimensions.

(1) *semantic accuracy*, refers to the degree to which data corresponds to real facts [35], i.e. the correctness of the available data. Although the open character of linked open data sources ensures a comparably high number of end users and inspires systematic studies on the data quality of the published datasets [8, 23, 27, 34], semantic accuracy is still an issue, given the vastness of the available data. Figure 2 visualizes an example, where the DBpedia entry for the Swiss ice hockey player Fritz Kraatz contains the incorrect `foaf:name` and `dbp:name` value "Barack Obama". Automatically addressing such issues is challenging, as the second snippet in Figure 2 illustrates. Although the DBpedia entry for the wrestler "The Sheik II" contains the `foaf:name` "NASDAQ" a subsequent investigation concludes that this name is actually used by the described individual and, therefore, considered valid. Algorithmic data assessment addresses these issues by combining data from heterogeneous data sources, applying additional background knowledge from dictionaries and using advanced disambiguation methods such as graph disambiguation (Section 4.1). Graph disambiguation selects relevant information based on the set of entities mentioned in a document, effectively mitigating semantic accuracy issues in linked data.

(2) *consistency* ensures that knowledge sources do not contain conflicting statements. An example for a consistency problem which is particularly relevant to named entity linking is the typing of resources. Although it is quite common for DBpedia entities to have multiple types such as `dbo:Person`, `yago:Athlete` and `yago:Actor`, the overlapping of types that exclude each other such as person, organization and location is troublesome. For instance, the current DBpedia version contains 4,606 entities that have both person *and* organization types assigned to them (Table 1).

(3) *conciseness* refers to the minimization of redundancy at the schema and data level. Conciseness at the schema level ensures that the dataset does not contain redundant properties and types. Conciseness at the data level, in contrast, prevents redundant entities. DBpedia, for instance, contains properties (e.g. `dbp:birthPlace` versus `dbo:birthPlace`)

Table 1: Entities may have multiple, conflicting entity types. The analyzed data contains a total of 3,147,225 named entities with any of those types.

	PERS	ORG	GEO
GEO	1,072	27,849	1,044,966
ORG	4,606	375,368	-
PERS	1,693,364	-	-

as well as types (`foaf:Person`, `dbo:Person`, `yago:Person-100007846`) with similar semantics. Another related problem are different data types and data semantics used with these properties. Some DBpedia entities, for example, specify the birth place with a literal, others with a URL to the corresponding location. Depending on the entity, a query will also retrieve only one location, or multiple locations indicating the birth place at different granularity levels (country, state, city, etc.). Information Extraction methods can mitigate conciseness issues at the schema level by performing more complex graph mining queries that consider all relevant properties and types, rather than a single one. Conciseness issues on the data level are especially common for organization, where a sensible separation between the organization, subsidiaries, branch offices and other legal entities is hard to accomplish. Weichselbraun et al. [31], for instance, report that large Swiss banks have been represented with 83 (Credit Suisse) and 92 (UBS) legal entities in a linked enterprise dataset. Graph-disambiguation and the use of weights based on an organization's page rank, turnover and number of employees are feasible strategies for selecting the most relevant of multiple redundant entities.

(4) *completeness* comprises multiple dimensions where the most relevant for information extraction are (i) property completeness, i.e. that all properties relevant for the information extraction task are available within the dataset, and (ii) population completeness which refers to the coverage in terms of individuals of the dataset. Effective mitigation strategies for a lack in completeness are algorithmic approaches for computing name variants (i.e. integrating the available data and context information to create candidates for missing values), and combining multiple data sources that complement each other. In cases, where this is not sufficient, researchers might even consider expanding publicly available data with custom linked data which encodes knowledge that is specific to the application domain.

Another key question related to completeness is, whether linked data is *required* to obtain usable results or whether it is merely a means for improving them. In the later case, researcher might implement mechanisms for handling missing data such as default values and fallback algorithms. Finally, the data interpretation and visualization step may distinguish between information covered in linked data sources and information where no background knowledge is available. The use case discussed in Section 5, for instance, provides richer visualizations and automatic grouping for images that refer to DBpedia entities

4 METHOD

This section discusses how knowledge extraction components can efficiently counter the data quality issues listed in Section 3. At first it introduces Recognyze, a component that supports data quality mitigation strategies to fully utilizes linked data for grounding mentions of entities in input documents to their respective resource in the underlying data sources. Afterwards, we describe the graph mining component used to enrich the named entities detected by Recognyze with background knowledge extracted from linked data repositories. Section 5 then demonstrates how the annotations provided by these components can be used to enable semantic search and browsing.

4.1 Named Entity Linking

Recognyze [31] is a named entity linking component that tightly integrates with linked data repositories. Its entity linking profiles specify (i) the data repositories to query, (ii) SPARQL statements for retrieving information from these repositories, (iii) a background knowledge acquisition pipeline that considers different kinds of filters, preprocessors and analyzers, as well as (iv) disambiguation algorithms for the task of identifying and grounding named entity mentions. To improve precision and recall, Recognyze's background knowledge acquisition pipeline aims on maximizing the extend and the quality of information received from repositories by drawing upon methods to automatically generate and validate name variants, and to analyze the background knowledge that describes the context of an entity.

Figure 3 illustrates these steps which aim on maximizing coverage and quality of the knowledge extracted from the underlying linked data sources. Preprocessors aim at generating and cleaning input data, analyzers assess these data and classify them into name, ambiguous name or context relevant to a particular named entity. Filters remove unwanted resources and misleading name variants upfront and after the preprocessing step. The extracted name variants, context information, relations and links between entities are used as background knowledge by the named entity linking component. A comprehensive library of filter, preprocessor and analyzer types, enables the adaption of Recognyze's background knowledge acquisition pipeline to different use cases and applications.

4.1.1 *Data completeness and conciseness.*

Recognyze has two build in mechanisms for addressing data sparsity: (i) it supports graph mining to counter data quality issues caused by *conciseness issues* and provides means to combine multiple linked data sources into a profile to address *completeness issues* within datasources. For instance, entities obtained from DBpedia can be enriched with background knowledge from Wikidata to obtain additional name variants for named entities (Figure 4), or with data from GeoNames to enrich DBpedia locations with hierarchical information (country, state, district, etc.). Cases, where the data coverage is not sufficient, can be tackled by creating and integrating individual datasets.

(ii) Recognyze's preprocessing component also enables the generation of *algorithmic name variance* by splitting an input string s into

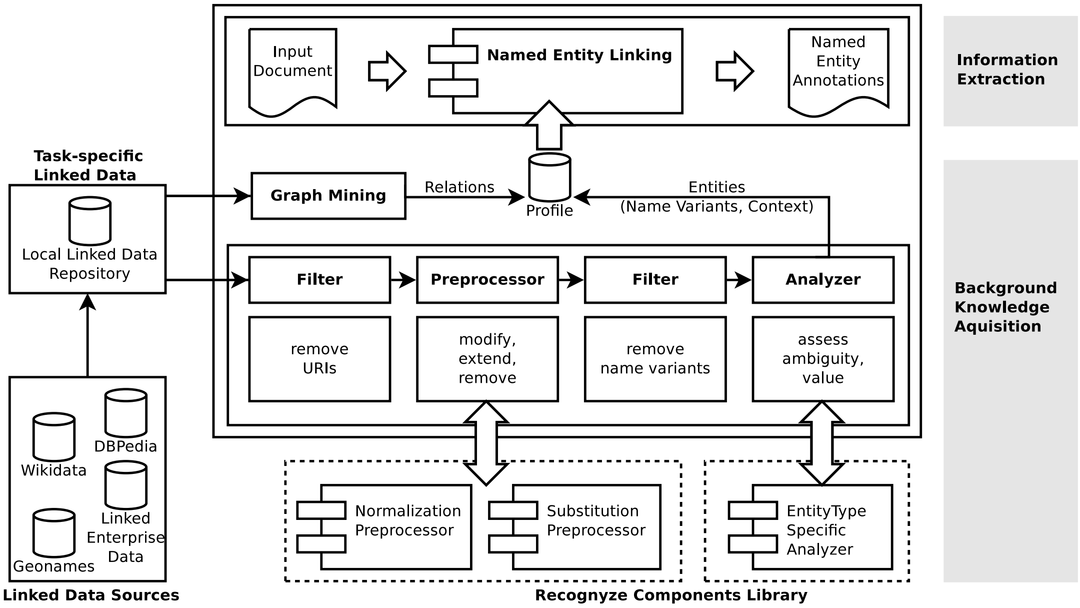


Figure 3: Recognyze's linked data driven background knowledge acquisition process.

tokens $t_i = \{t_1, ...t_n\}$ that are then used to generate name variants $n^1, ...n^m$. A simple variance of this algorithm returns substrings $n^1 = t_1$; $n^2 = t_1 t_2$; ... $n^{n-1} = t_1 t_2 ... t_{n-1}$ and replaces tokens t_i with synonyms, if available. For instance, the name "United States Department of Commerce" yields the additional name variants "U.S. Department of Commerce" and "US Department of Commerce" in addition to the substring variances of these strings. The name variant algorithm also considers heuristics for handling uppercase names and automatically generates name variants from names that use non ASCII characters by drawing upon Unicode's Normalization Form Canonical Decomposition (NFD).

```
### DBpedia
SELECT DISTINCT ?s ?name WHERE {
  { ?s rdfs:label ?name.
    ?s a dbo:Place. }
  UNION
  { ?s rdfs:label ?name.
    ?s a dbo:Location. }
}
### Wikidata
SELECT DISTINCT ?s ?alternativename WHERE {
    ?wikidata owl:sameAs ?s.
    ?wikidata rdfs:label ?alternativename.
    FILTER(lang(?alternativename) = "en")
    FILTER(regex(str(?s), "dbpedia")) }
```

Figure 4: Example queries for a simple profile that combines and enriches DBpedia entities (top) with alternative names obtained from Wikidata (bottom).

4.1.2 Semantic accuracy and consistency. Recognyze uses preprocessors, filters and analyzers to improve data quality in terms of *semantic accuracy*. Depending on the use case, preprocessors might require a certain string length, filter names that do not contain any letters, or remove invalid characters from names.

Analyzers use more complex algorithms such as entropy metrics [31] for evaluating name variants and to judge whether such a variance is considered unique enough to refer to a mention of a named entity. For ambiguous entities Recognyze uses prefixes and suffixes that are specific to the entity type (e.g. Dr, Prof, President, etc. for people) to assess whether a particular mention refers to that entity type.

Recognyze does not yet provide automated ways for dealing with *data consistency issues* such as incompatible and overlapping entity types. The entity `dbr:United_Daughters_of_the_Confederacy` that is labelled as both, an organization and a person, for example, is currently processed twice, once for each entity type, leading to potential problems in the named entity linking step.

Options for mitigating this kind of *consistency* problems are:

- cleaning the task-specific linked data source, or using blacklist filtering to remove the problematic entity types. Since these are manual steps, they are usually time intensive, costly and not feasible for large knowledge repositories.
- automatic resolution of conflicting values by establishing a preference order. This strategy basically assigns the type with the highest preference to the named entity. The preference order is hereby determined by analyzing a random sample of overlapping types. While this will work in many cases, it also introduces a potential error source into the named entity linking process.
- Implementing conflict resolution is the most complex but also most reliable solution. Strategies for resolving type

Table 2: Simple example queries for the data enrichment service.

entity type	query	use case (example ← selected values)

location	**SELECT** ?lat ?long ?population ?country
	WHERE {
	?e wgs84_pos:lat ?lat .
	?e wgs84_pos:long ?long .
	OPTIONAL
	{ ?e gn:population ?population .}
	OPTIONAL
	{ ?e gn:parentCountry ?country . } }
		Mark the location on a map and provide metadata such as population and country.
		(Switzerland ← lat: 46.84986, long: 9.53287, population: 32429, country: geo:2661169)

organization	**SELECT** ?key_person **WHERE** {
	{ ?e dbo:keyPerson ?key_person .}
	UNION
	{ ?e dbp:keyPeople ?key_person .}
	UNION
	{ ?e dbo:occupation ?key_person .} }
		An organization's key people.
		(Apple Inc. ← dbr:Arthur_D._Levinson, dbr:Jonathan_Ive, dbr:Luca_Maestri, dbr:Tim_Cook)

person	**SELECT** ?type **WHERE** {
	?e rdf:type ?type }
		Automatic image grouping.
		(Barack Obama ← yago:HeadOfState, yago:-President110467179, yago:StateSenator110650076)

conflicts could use graph mining algorithms or draw upon the entity's properties to compute the most likely entity type. Locations, for instance, have a significantly higher probability of having a "geo:lat" property than a person or organization.

Newer version of Recognyze also consider graph disambiguation which leverages links between named entities within the knowledge source for disambiguation.

4.2 Data Enrichment

A graph mining and data enrichment component focuses on obtaining background information on the extracted named entities which is instrumental for (i) providing additional features for subsequent data mining tasks, (ii) enriching named entities with properties and values that might be useful in later reasoning and data cleanup steps which identify invalid values, violated constraints, syntactic errors, etc., (iii) mining knowledge that supports search, browsing and visualizing the extracted information.

Figure 5 illustrates the graph mining and data enrichment component. The component obtains named entities from different input formats ranging from simple named entity lists to more structured formats such as the *NLP Interchange Format (NIF)*. The graph mining draws upon configuration profiles that specify SPARQL queries, the corresponding repositories and relevant metadata fields to mine the local linked data repository for relevant knowledge, and an entity formatter exports the enriched named entities to a user-specified output format.

Table 2 illustrates a number of example queries and their application for the use case presented in this paper. The location query, for instance, obtains data relevant for displaying the location on a map with metadata on its population and geo hierarchy. The example for organizations obtains key people from DBpedia that can be used to provide context information for the entity. The person query, in contrast, yields the entity's type, enabling the automatic grouping of entities (e.g. politicians, athletes, artists, ..).

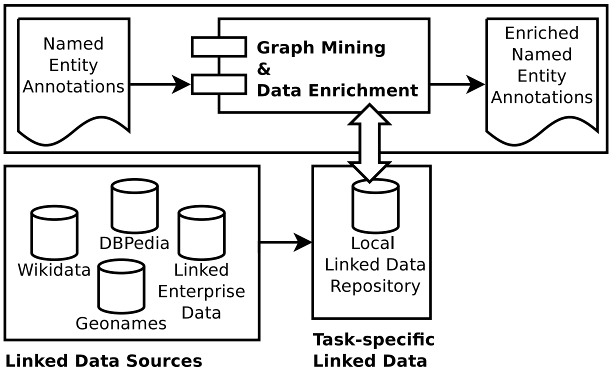
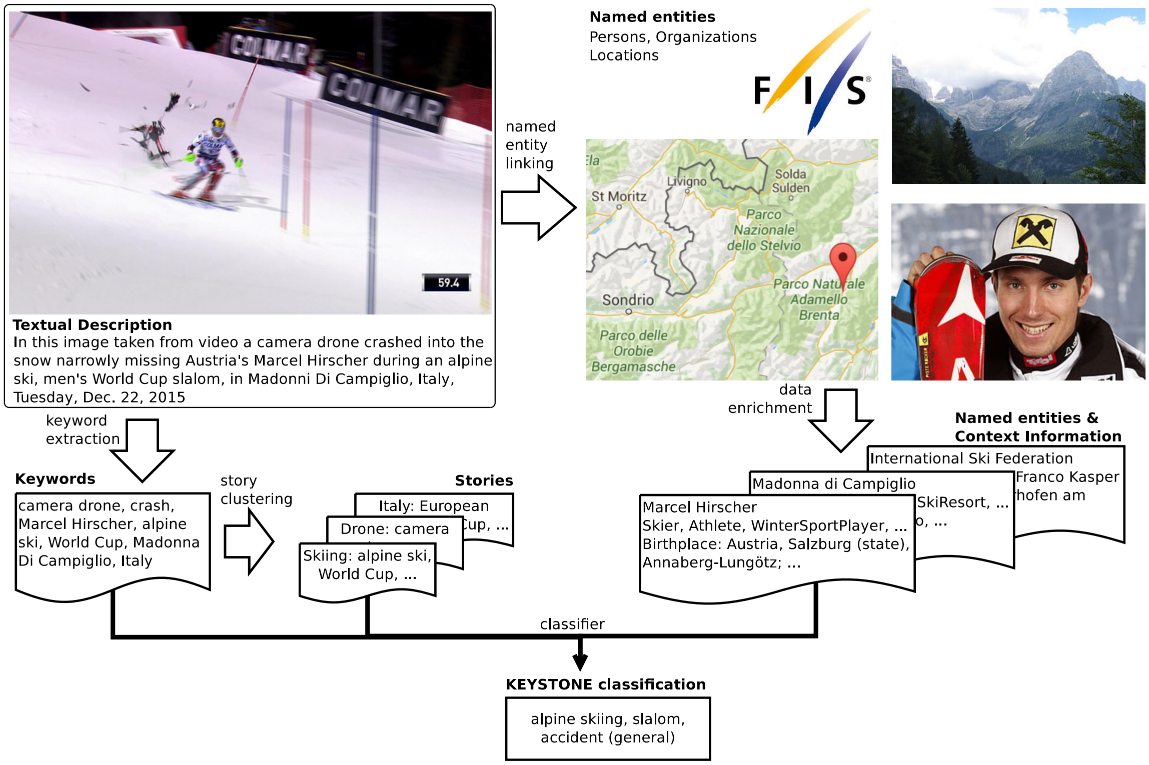

Named Entity Annotations
Graph Mining & Data Enrichment
Enriched Named Entity Annotations
Wikidata
DBPedia
Geonames
Linked Enterprise Data
Local Linked Data Repository
Linked Data Sources
Task-specific Linked Data

Figure 5: Graph Mining and Data Enrichment.

5 USE CASE: CROSS-MODAL INFORMATION EXTRACTION FOR IMPROVED IMAGE META DATA (IMAGINE)

The retrieval and marketability of visual content highly depends on the availability of high quality image meta data which enables customers to efficiently locate relevant content in large image collections. Within the IMAGINE project we collaborate with KEYSTONE, Switzerland's largest provider of visual content, to develop advanced methods which exploit the convergence between textual image descriptions and image content as well as linked open data for information extraction and to automatically obtain relevant meta data such as keywords, named entities and topics.



Textual Description
In this image taken from video a camera drone crashed into the snow narrowly missing Austria's Marcel Hirscher during an alpine ski, men's World Cup slalom, in Madonni Di Campiglio, Italy, Tuesday, Dec. 22, 2015

59.4
named entity linking
Named entities
Persons, Organizations
Locations
FIS
St Moritz
Livigno
Solda
Sulden
Parco Nazionale dello Stelvio
Sondrio
Parco Naturale Adamello Brenta
Parco delle Orobie Bergamasche
data enrichment
keyword extraction
Keywords
camera drone, crash, Marcel Hirscher, alpine ski, World Cup, Madonna Di Campiglio, Italy
story clustering
Stories
Italy: European Cup, ...
Drone: camera
Skiing: alpine ski, World Cup, ...
Named entities & Context Information
International Ski Federation
Franco Kasper
hofen am
Madonna di Campiglio
SkiResort, ...
o, ...
Marcel Hirscher
Skier, Athlete, WinterSportPlayer, ...
Birthplace: Austria, Salzburg (state),
Annaberg-Lungötz; ...
classifier
KEYSTONE classification
alpine skiing, slalom, accident (general)

Figure 6: IMAGINE automatic image annotation pipeline.

The presented use case addresses the following three key challenges within KEYSTONE's image annotation process: (i) the trade-off between cost and throughput, (ii) keyword spamming, and (iii) metadata quality and consistency.

The *trade-off between cost and throughput* is a major issue controlling meta data quality. KEYSTONE employs a team of domain experts who manually enrich photographs with image metadata and assign images to a controlled vocabulary of topics (broad terms) and keywords (specific descriptors) within the KEYSTONE ontology which is a customized version of the IPTC photo metadata standard (iptc.org/standards/photo-metadata/iptc-standard). Currently, these experts annotate between 300 and 500 photographs per day that have been contributed by photographers distributing their photographs directly through KEYSTONE. In addition, KEYSTONE also receives an image stream of approximately 20,000 photographs per day from third party agencies, where the company is forced to rely on the embedded metadata since it is not economically feasible to manually annotate these image streams.

These third-party images are also prone to *metadata spamming*, i.e. the practice of adding frequently used search terms such as *dog* and *cat* to images, regardless of their actual content. These additional tags increase an image's likelihood to appear in search results and, therefore, are very attractive to the third party agency but not to KEYSTONE and its customers since bogus annotations reduce the effectiveness of the image search and lead to irrelevant results.

Finally, *metadata quality and consistency* are also an issue. Although metadata created by KEYSTONE's domain experts shows a considerably higher quality, investigations within the IMAGINE project uncovered that experts tend to use only a fraction of the available vocabulary. Therefore, parts of the controlled vocabulary are only used barely at all.

The IMAGINE project develops advanced information extraction methods to address these issues by (i) providing the option to automate metadata generation for third party image streams and to support manual annotation processes by computing suggestions for the domain experts. The computer suggestions draw upon mandatory textual image descriptions that are (ii) considerably harder to manipulate since statistical natural language processing methods automatically reduce the significance of terms that are overrepresented within these descriptions. Finally, (iii) these methods are expected to provide high quality metadata as long as human experts ensure that the system receives sufficient training data and feedback.

The following sections elaborate on the application of the methods described in Section 4 to the presented use case, discuss the impact of linked data quality on the developed components, and describe the chosen mitigation strategies.

5.1 Image Annotation Process

Figure 6 outlines the automatic image annotation process developed within the IMAGINE project. IMAGINE draws upon mandatory textual image descriptions, that are present in all visual content regardless of whether it has been created by contract photographers or by third-party agencies. As described in Section 4.1, named entity linking then identifies mentions of persons, organizations and locations based on background knowledge obtained from DBpedia, Wikidata and GeoNames dumps that have been loaded into a task-specific linked data repository. Using these entities as seed terms for a subsequent graph mining and data enrichment process (Section 4.2) yields information on the extracted named entities. Graph mining allows enriching

(1) persons such as Marcel Hirscher with entity classes (`rdf:type`) – e.g. Skier, Athlete, Austrian Alpine Skier) and events in which these persons participated (FIS Alpine World Ski Championships, Winter Olympics, etc.),
(2) locations with coordinates and elevation (`georss:point`, `dbo:maximumElevation`, `dbo:minimumElevation`) and entity classes (Ski Area, Ski Resort), and
(3) organizations with information on events they organized (`dbo:organised`) – e.g. Four Hills Tournament, FIS Nordic World Ski Championship, etc. – links between organizations (`dbp:parentOrganization`, `dbp:membership`, `dbp:-affof`, and key personnel within the organization (`dbo:keyPerson`, `dbp:leaderName`).

In addition, DBpedia provides thumbnails for many entries that can be used to group items in search queries.

The named entity linking and data enrichment process provides the following main benefits to KEYSTONE and its customers: (i) it paves the way for semantic search capabilities, i.e. locating images that contain linked named entities not only by the terms used in the image description but also by alternative names and types such as *Alpine Skier* and *Athlete*, (ii) provides additional dimensions for faceted browsing and limiting search results, and (iii) it facilitates the automatic generation of image collections based on common properties (people participating in a certain event, associated with organizations or political parties, photographs covering a certain region, etc.).

As visualized in Figure 6, the annotation pipeline also employs keyword extraction [30] to extract descriptor terms (k_i). Transforming these descriptors into an undirected graph $G = (V, E)$ with

(1) keywords $k_i \in V$ as vertices V and
(2) edges E between pairs of keywords (k_i, k_j) that appear in the same image description d

enables using the Louvaine algorithm for fast community detection [2] to perform story clustering based on the extracted keywords.

In the final stage, a classification component integrates (i) identified named entities, (ii) the background information obtained by the data enrichment process, (iii) the extracted keywords, and (iv) clustered stories to assign images to the controlled vocabulary within the KEYSTONE ontology.

5.2 Discussion

The presented system provides KEYSTONE with effective means to support its own annotation process (e.g. by providing suggestions to its domain experts). In addition, using the pipeline without supervision provides an cost-effective way to improve the metadata quality in third-party images.

Table 3: Number of persons (PER), organizations (ORG) and locations (LOC) in the DBpedia 201510 dump and the online version (March 2017) depending on the query.

entity	properties	number of matches	
		201510	online
PER	a foaf:Person	1,588,591	1,518,282
	a dbo:Person	1,587,944	1,517,816
	a yago:Person100007846	1,010,473	1,216,440
	UNION	1,699,042	1,716,717
ORG	a dbo:Organisation	267,318	275,077
	a yago:Organization108008335	280,243	333,269
	a dbo:Company	65,754	67,544
	UNION	407,823	426,445
LOC	a dbo:Place	801,334	816,252
	a dbo:Location	801,334	816,252
	a yago:YagoGeoEntity	814,258	989,272
	geo:lat	970,143	970,143
	UNION	1,111,634	1,217,453

Named entity linking and the corresponding semantic enrichment also pave the way for a number of new applications such as semantic search, automatic image grouping and the suggestion of related images based on concepts and links obtained from DBpedia.

Although the potential of the semantic enrichment is impressive, our experiments also show that data quality, especially in terms of *completeness* but also in terms of *conciseness* and *semantic accuracy* poses a serious limitation on the data's usefulness.

Table 3 outlines the number of persons, organizations and locations obtained from the DBpedia SPARQL endpoint in March 2017. Since there is no single concise way to query for these entity types, the chosen query has a significant impact on the returned results. As outlined in Section 3 this problem can be addressed by graph mining - i.e. running (more) complex queries on the datasource that capture a higher fraction of relevant entities.

Table 4 provides another example of how more complex queries can address *conciseness* and *completeness* issues.

The table lists the number of medals various countries have won at the 2014 Winter Olympics in Sotschi comparing the official statistics (Real World) with results obtained from DBpedia in March 2017. One interesting observation indicating data conciseness issues is that the number of medals returned depends on the used query - i.e. whether dbp:birthPlace or dbo:birthPlace are used to identify an athlete's nationality and which class is used in the rdf:type query.

Comparing the DBpredia results to the official medal statistics (Real World) yields another interesting insight. The number of medals returned for some countries exceeds their official medal count. Canada, for instance, received 25 medals in the games, but DBpedia lists up to 51 athletes that received medals which would suggest a *semantic accuracy* issue. Querying the wining athletes shows that this discrepancy is caused by team sports such as ice

Table 4: Medalists at the 2014 Winter Olympics in Sotschi.

source	query	total medals	top three countries
Real World	-	295	Russia (33) United States (28) Norway (26)
DBpedia	yago:WikicatMedalistAtThe2014WinterOlympics dbp:birthPlace	268	Canada (49) Sweden (31) Finland (29)
DBpedia	yago:WikicatMedalistAtThe2014WinterOlympics dbo:birthPlace	272	Canada (51) Sweden (32) Finland (29)
DBpedia	dbc:Alpine_skiers_at_the_2014_Winter_Olympics dbp:birthPlace	426	Canada (49) Sweden (36) Finland (29)
DBpedia	dbc:Alpine_skiers_at_the_2014_Winter_Olympics dbo:birthPlace	389	Canada (51) Sweden (37) Finland (29)

hockey, where all of the team members would be marked as medal winning athletes. Actually, 47 of the listed athletes for Canada are ice hockey players which indicates that at most five of the 25 disciplines in which the country received medals are listed in DBpedia.

Addressing *conciseness* issues by combining attributes improves the *completeness* of the returned results at the cost of complexity.

The presented example also demonstrates the importance of *timeliness* in the data. The analysis covered the Winter Games in Sotschi since data from the Olympic Summer Games in Rio hadn’t been published in DBpedia by March 2017. Considering such restrictions in the selection process, where the usefulness of data sources for particular tasks and use cases is evaluated, is essential.

In conclusion, industrial applications that draw upon DBpedia need to be very conscious about limitations in the data quality and whether these imitations have a serious impact on the usefulness of the developed systems. In the present use case, for example, we had to choose an *adaptive method design* that makes background knowledge from linked data optional for both semantic search and faceted browsing since the availability of such knowledge for all entities cannot be guaranteed.

Nevertheless these data is still useful for both tasks. Semantic search and the automatic generation of image collections benefit from the retrieved background knowledge, although it is not available for all entities. Faceted browsing, in contrast, is more seriously affected by missing metadata, since it may lead to empty results for selected facets that are caused by missing data in the knowledge source rather than missing photographs. Fallback mechanism such as providing a default category for entities with no background knowledge are remedies to mitigate this problem.

6 CONCLUSIONS

This paper described linked data quality issues and mitigation strategies that help in using linked data as knowledge source in knowledge-intensive information extraction processes. One major insight gained is that useful data does not necessarily correspond to data without any quality issues. Similar to the World Wide Web the scope and extend of the available data plays a major role in its usability. Prime examples for vast data sources are DBpedia, GeoNames and Wikidata which are highly attractive for information extraction, despite the considerable number of quality issues that can be observed in these datasets. The ability to mitigate data quality issues is, therefore, key to successfully retrieving and exploiting background knowledge from these sources for information extraction tasks. The linked data quality dimensions and mitigation strategies presented in this paper aim at supporting researchers in understanding and addressing linked data quality issues, and to inspire research on the design of knowledge-rich information extraction methods.

Future work will focus on (i) advancing research on mitigation strategies and (ii) investigating means to address the problem of domain and knowledge-base evolution and its interaction with information extraction methods. For instance, if a mention of “U.S. president” is grounded against DBpedia, the grounding depends on the chosen knowledge base version. Identifying consistent strategies for versioning knowledge bases and information extraction artifacts that have been created with a certain knowledge base version is an important cornerstone for a reliable handling of knowledge evolution and other temporal effects relevant to information extraction.

Acknowledgment

The research presented in this paper has been conducted as part of the IMAGINE Project (www.htwchur.ch/imagine), funded by the Swiss Commission for Technology and Innovation (CTI). The authors would like to thank the KEYSTONE AG and especially Markus Läng, Karl Csoknyay, Sven Rütti and Lucas Kobel for their valuable input in designing the image annotation and retrieval processes.

REFERENCES

[1] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.. In *International Conference on Language Resources and Evaluation (LREC 2010)*. Malta, 2200–2204.

[2] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (Oct. 2008).

[3] Erik Cambria, Soujanya Poria, Rajiv Bajpai, and Björn W. Schuller. 2016. SenticNet 4: A Semantic Resource for Sentiment Analysis Based on Conceptual Primitives. In *26th International Conference on Computational Linguistics (COLING 2016)*. 2666–2677.

[4] Erik Cambria and Bebo White. 2014. Jumping NLP Curves: A Review of Natural Language Processing Research. *IEEE Computational Intelligence Magazine* 9, 2 (May 2014), 48–57.

[5] Rodolfo C. Cavalcante, Rodrigo C. Brasileiro, Victor L. F. Souza, Jarley P. Nobrega, and Adriano L. I. Oliveira. 2016. Computational Intelligence and Financial Markets: A Survey and Future Directions. *Expert Systems with Applications* (2016).

[6] Wingyan Chung and Daniel Zeng. 2016. Social-media-based public policy informatics: Sentiment and network analyses of U.S. Immigration and border security. *Journal of the Association for Information Science and Technology* 67, 7 (2016), 1588–1606.

[7] Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. 2013. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-SEMANTICS'13)*. 121–124.

[8] Jeremy Debattista, Soeren Auer, and Christoph Lange. 2016. Luzzu - A Framework for Linked Data Quality Assessment. In *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)*. IEEE; IEEE Comp Soc, IEEE, 124–131.

[9] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. Advances in Knowledge Discovery and Data Mining. American Association for Artificial Intelligence, 1–34.

[10] Xianpei Han and Jun Zhao. 2009. Named entity disambiguation by leveraging wikipedia semantic knowledge. In *Proceedings of the 18th ACM conference on Information and knowledge management (CIKM '09)*. ACM, 215–224.

[11] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*. Association for Computational Linguistics, 782–792.

[12] Saurabh S. Kataria, Krishnan S. Kumar, Rajeev R. Rastogi, Prithviraj Sen, and Srinivasan H. Sengamedu. 2011. Entity disambiguation with hierarchical topic models. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '11)*. ACM, 1037–1045.

[13] Erin Hea-Jin Kim, Yoo Kyung Jeong, Yuyoung Kim, Keun Young Kang, and Min Song. 2015. Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news. *Journal of Information Science* (Oct. 2015).

[14] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2014. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal* (2014).

[15] Qing Li, TieJun Wang, Ping Li, Ling Liu, Qixu Gong, and Yuanzhu Chen. 2014. The effect of news and public mood on stock movements. *Information Sciences* 278 (Sept. 2014), 826–840.

[16] Nandana Mihindukulasooriya, Mariano Rico, Raul Garcia-Castro, and Asuncion Gomez-Perez. 2015. An Analysis of the Quality Issues of the Properties Available in the Spanish DBpedia. In *Advances in Artificial Intelligence (CAEPIA 2015)*. 198–209. 16th Conference of the Spanish-Association-for-Artificial-Intelligence (CAEPIA), Albacete, SPAIN, NOV 09-12, 2015.

[17] Dat Ba Nguyen, Johannes Hoffart, Martin Theobald, and Gerhard Weikum. 2014. AIDA-light: High-Throughput Named-Entity Disambiguation. In *Linked Data on the Web, WWW 2014*, Vol. 1184. Seoul, South Korea.

[18] Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence* 194 (2013), 151–175.

[19] Heiko Paulheim and Christian Bizer. 2014. Improving the Quality of Linked Data Using Statistical Distributions. *International Journal on Semantic Web and Information Systems* 10, 2 (2014), 63–86.

[20] Anja Pilz and Gerhard Paaß. 2011. From names to entities using thematic context distance. In *Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM '11)*. ACM, 857–866.

[21] Suhas Ranganath, Xia Hu, Jiliang Tang, and Huan Liu. 2016. Understanding and Identifying Advocates for Political Campaigns on Social Media. In *Proceedings of the 9th ACM International Conference on Web Search and Data Mining (WSDM'16)*.

[22] Petar Ristoski and Heiko Paulheim. 2016. Semantic Web in data mining and knowledge discovery: A comprehensive survey. *Web Semantics: Science, Services and Agents on the World Wide Web* 36 (Jan. 2016), 1–22.

[23] Edna Ruckhaus, Maria-Esther Vidal, Simon Castillo, Oscar Burguillos, and Oriana Baldizan. 2014. Analyzing Linked Data Quality with LiQuate. In *Semantic Web: ESWC 2014 Satellite Events*. 488–493.

[24] Arno Scharl and David D. Herring. 2013. Extracting Knowledge from the Web and Social Media for Progress Monitoring in Public Outreach and Science Communication. In *Proceedings of the 19th Brazilian Symposium on Multimedia and the Web (WebMedia '13)*. ACM, 121–124.

[25] Arno Scharl, Albert Weichselbraun, Max Göbel, Walter Rafelsberger, and Ruslan Kamolov. 2016. Scalable Knowledge Extraction and Visualization for Web Intelligence. In *Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS-49)*. IEEE Computer Society Press.

[26] Robert Speer and Catherine Havasi. 2012. Representing General Relational Knowledge in ConceptNet 5. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA).

[27] Harsh Thakkar, Kemele M. Endris, Jose M. Gimenez-Garcia, Jeremy Debattista, Christoph Lange, and Sören Auer. 2016. Are Linked Datasets Fit for Open-domain Question Answering? A Quality Assessment. In *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics (WIMS '16)*. ACM, New York, NY, USA, Article 19, 12 pages.

[28] Albert Weichselbraun, Stefan Gindl, Fabian Fischer, Svitlana Vakulenko, and Arno Scharl. 2016. Aspect-Based Extraction and Analysis of Affective Knowledge from Social Media Streams. *IEEE Intelligent Systems* (2016). Accepted 30 June 2016.

[29] Albert Weichselbraun, Stefan Gindl, and Arno Scharl. 2013. Extracting and Grounding Context-Aware Sentiment Lexicons. *IEEE Intelligent Systems* 28, 2 (2013), 39–46.

[30] Albert Weichselbraun, Arno Scharl, and Stefan Gindl. 2016. Extracting Opinion Targets from Environmental Web Coverage and Social Media Streams. In *Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS-49)*. IEEE Computer Society Press.

[31] Albert Weichselbraun, Daniel Streiff, and Arno Scharl. 2015. Consolidating Heterogeneous Enterprise Data for Named Entity Linking and Web Intelligence. *International Journal on Artificial Intelligence Tools* 24, 2 (2015).

[32] Shengsheng Xiao, Chih-Ping Wei, and Ming Dong. 2016. Crowd intelligence: Analyzing online product reviews for preference measurement. *Information & Management* 53, 2 (March 2016), 169–182.

[33] Donghui Yang, Chao Huang, and Mingyang Wang. 2016. A social recommender system by combining social network and sentiment similarity: A case study of healthcare. *Journal of Information Science* (2016).

[34] Amrapali Zaveri, Dimitris Kontokostas, Mohamed A. Sherif, Lorenz Bühmann, Mohamed Morsey, Sören Auer, and Jens Lehmann. 2013. User-driven Quality Evaluation of DBpedia. In *Proceedings of the 9th International Conference on Semantic Systems (I-SEMANTICS '13)*. ACM, 97–104.

[35] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. 2016. Quality assessment for Linked Data: A Survey. *Semantic Web* 7, 1 (Jan. 2016), 63–93.