Retrieval, Crawling and Fusion of Entity-centric Data on the Web

Stefan Dietze

L3S Research Center, Leibniz Universität Hannover, Germany
dietze@L3S.de

Abstract. While the Web of (entity-centric) data has seen tremendous growth over the past years, take-up and re-use is still limited. Data vary heavily with respect to their scale, quality, coverage or dynamics, what poses challenges for tasks such as entity retrieval or search. This chapter provides an overview of approaches to deal with the increasing heterogeneity of Web data. On the one hand, recommendation, linking, profiling and retrieval can provide efficient means to enable discovery and search of entity-centric data, specifically when dealing with traditional knowledge graphs and linked data. On the other hand, embedded markup such as Microdata and RDFa has emerged a novel, Web-scale source of entity-centric knowledge. While markup has seen increasing adoption over the last few years, driven by initiatives such as schema.org, it constitutes an increasingly important source of entity-centric data on the Web, being in the same order of magnitude as the Web itself with regards to dynamics and scale. To this end, markup data lends itself as a data source for aiding tasks such as knowledge base augmentation, where data fusion techniques are required to address the inherent characteristics of markup data, such as its redundancy, heterogeneity and lack of links. Future directions are concerned with the exploitation of the complementary nature of markup data and traditional knowledge graphs.

Keywords: Entity Retrieval, Dataset Recommendation, Markup, schema.org, Knowledge Graphs

1 Introduction

The emergence and wide-spread use of knowledge graphs, such as Freebase [5], YAGO [31], or DBpedia [1] as well as publicly available linked data [2], has led to an abundance of entity-centric available on the Web. Data is shared as part of datasets, usually containing interdataset links [23], which link equivalent, similar or related entities, while the majority of these links are concentrated on established reference graphs [12]. Datasets vary significantly with respect to represented resource types, currentness, coverage of topics and domains, size, used languages, coherence, accessibility [7] or general quality aspects [16].

Also, while entity-centric knowledge bases capture large amounts of factual knowledge in the form of RDF triples (subject-predicate-object), they still are incomplete and inconsistent [39], i.e., coverage, quality and completeness vary heavily across types or domains, where in particular long-tail entities usually

are insufficiently represented. In addition, while sharing of vocabularies and vocabulary terms is a crucial requirement for enabling reuse, the Web of data still features a large amount of highly overlapping and often unlinked vocabularies [8].

The wide variety and heterogeneity of available data(sets) and their characteristics pose significant challenges for data consumers when attempting to find and reuse useful data without prior knowledge about the available data and their features. This is seen as one of the reasons for the strong bias towards reusing well-understood reference graphs like Freebase or Yago, while there exists a long tail of datasets which is hardly considered or reused by data consumers.

In this work, we discuss a range of research results which aim at improving search and retrieval of entity-centric Web data. In Section 2, we will focus specifically on approaches towards dealing with the aforementioned heterogeneity of available knowledge graphs and linked data by introducing methods for recommendation and profiling of datasets as well as for enabling efficient entity retrieval. In Section 3, we will look beyond traditional linked data and discuss new forms of emerging entity-centric data on the Web, namely structured Web markup annotations embedded in HTML pages. Markup data has become prevalent on the Web, building on standards such as RDFa[1], Microdata[2] and Microformats[3], and driven by initiatives such as schema.org, a joint effort led by Google, Yahoo!, Bing and Yandex. We will introduce a number of case studies about scope and coverage of Web markup and introduce recent research which aims at exploiting Web markup data for tasks such as knowledge base population, data fusion or entity retrieval.

While this paper aims at providing a subjective overview of recent works as well as current and future research directions for the exploitation of entity-centric Web data, it is worthwhile to highlight that an exhaustive survey is beyond the scope of this paper.

2 Recommendation, Profiling and Retrieval of Entity-centric Web Data

The growth of structured linked data on the Web covers cross-domain and domain-specific data from a wide range of domains, where bibliographic (meta)data, such as [10], general resource metadata [32] and data from the life sciences domain [11] are particularly well represented. However, reuse of both vocabularies [8] as well as data is still limited. Particularly with respect to interlinking, the current topology of the linked data Web graph underlines the need for practical and efficient means to recommend suitable datasets: currently, only few, well established knowledge graphs show a high amount of inlinks, with DBpedia being the most obvious target [30], while a long tail of datasets is largely ignored.

[1] RDFa W3C recommendation: http://www.w3.org/TR/xhtml-rdfa-primer/
[2] http://www.w3.org/TR/microdata
[3] http://microformats.org

To facilitate search and reuse of existing datasets, descriptive and reliable metadata is required. However, as witnessed in the popular dataset registry DataHub[4], dataset descriptions often are missing entirely, or are outdated, for instance describing unresponsive endpoints [7]. This issue is partially due to the lack of automated mechanisms for generating reliable and up-to-date dataset metadata and hinders the reuse of datasets. The dynamics and frequent evolution of data further exacerbates this problem, calling for scalable and frequent update mechanisms of respective metadata.

In this section, we will introduce approaches aiming at facilitating data reuse and retrieval through (a) automated means for dataset recommendation, (b) dataset profiling as a means to facilitate dataset discovery through generating descriptive dataset metadata, and (c) improved entity retrieval techniques which address the heterogeneity of Web data, particularly, the prevalent lack of explicit links.

2.1 Dataset Recommendation

Dataset recommendation approaches such as [20] and [27] or the more recent works [12] and [13] tackle the problem of computing a ranking of datasets of relevance for the linking task, i.e. likely to contain linking candidates for a given source dataset. Formally speaking, dataset recommendation considers the problem of computing a rank score for each elements of a set of target datasets D_T so that the rank score indicates the relatedness of D_T to a given source dataset.

Leme *et al.* [19] present a ranking method based on Bayesian criteria and on the popularity of the datasets, what affects the applicability of the approach. The authors address this drawback in [20] by exploring the correlation between different sets of features, such as properties, classes and vocabularies.

Motivated by the observation that datasets often reuse vocabulary terms, [13] adopts the notion of a dataset profile, defined as a set of concept labels that describe the dataset and propose the *CCD-CosineRank* dataset recommendation approach, based on schema similarity across datasets. The approach consists of identifying clusters of comparable datasets, and, ranking the datasets in each cluster with respect to a given dataset. For the latter step, three different similarity measures are considered and evaluated. The approach is applied to the real-world datasets from the Linked Open Data graph and compared to two baseline methods, where results show a mean average precision of around 53% for recall of 100%, which indicates that *CCD-CosineRank* can reduce considerably the cost of dataset interlinking. As a by-product, the system returns sets of schema concept mappings between source and target datasets.

However, next to schema-level features, consideration of instance-level characteristics is crucial when computing overlap and complementarity of described entities. Given the scale of available datasets, exhaustive comparisons of schemas and instances or some of their features are not feasible as an online process. For instance, in [14] (Section 2.3), authors generate a weighted bipartite graph, where

[4] http://www.datahub.io

datasets and topics represent the nodes, related through weighted edges, indicating the relevance of a topic for a specific dataset. While computation of such topic profiles is costly, it is usually applied to a subset of existing datasets only, where any new or so far unannotated datasets require the pre-computation of a dedicated topic profile.

[12] builds on this observation and provides a recommendation method which not only takes into account the direct relatedness of datasets as emerging from the topic-dataset graph produced through the profiling in [14], but also adopts established collaborative filtering (CF) practices by considering the topic relationships emerging from the global topic-dataset-graph to derive specific dataset recommendations. CF enables to consider arbitrary (non-profiled) datasets as part of recommendations. This approach on the one hand significantly increases the recall of recommendations, and at the same time improves recommendations through considering dataset connectivity as another relatedness indicator. The intuition is that global topic connectivity provides reliable connectivity indicators even in cases where the underlying topic profiles might be noisy, i.e. that, even poor or incorrect topic annotations will serve as reliable relatedness indicator when shared among datasets. Theoretically, this approach is agnostic to the underlying topic index. This approach also reflects both, instance-level as well as schema-level characteristics of a specific dataset. Even though topics are derived from instances, resources of particular types show characteristic topic distributions, which significantly differ across different types [34].

Applied to the set of all available linked dataasets, experiments show superior performance compared to three simple baselines, namely based on shared key-words, shared topics, and shared common links, achieving a reduction of the original search space of up to 86% on average. It is worth to highlight that the aforementioned evaluation results are affected by the limited nature of available ground truth data, where all works relied on linkset descriptions from the DataHub. However, while this data is manually curated, it is inherently sparse and incomplete, that is, providers usually indidate a very limited amount of linking information. This leads to inflated recall values and at the same time, affects precision in the sense that results tend to label correct matches as false positives according to the ground truth. One future direction of research aims at producing a more complete ground truth. Given the scale of available data on the Web, computing linking metrics should resort to sampling and approximation strategies.

2.2 Dataset Profiling

Rather than automatically recommending datasets, additional metadata can enable data consumers to make an informed decision when selecting suitable datasets for a given task. In [14], authors address this challenge of automatically extracting dataset metadata with the goal of facilitating dataset search and reuse. Authors propose an approach for creating structured dataset profiles, where a profile describes the topic coverage of a particular dataset through a

weighted graph of selected DBpedia categories. The approach consists of a processing pipeline that combines tailored techniques for dataset sampling, topic extraction from reference datasets and topic relevance ranking. Topics are extracted through named entity recognition (NER) techniques and the use of a reference category vocabulary, namely DBpedia. Relevance of a particular category for a dataset is computed based on graphical models like *PageRank* [6], *K-Step Markov* [36], and *HITS* [18]. While this is a computationally expensive process, authors experimentally identify the parameters which enable a suitable trade-off between representativeness of generated profiles and scalability. Finally, generated dataset profiles are exposed as part of a public structured dataset catalog based on the *Vocabulary of Interlinked Datasets* (VoID[5]) and the recent vocabulary of links (VoL)[6].

As part of the experiments, authors generated dataset profiles for all accessible linked datasets classified as Linked Open Data on the DataHub and demonstrate that, even with comparably small sample sizes (10%), representative profiles and rankings can be generated. For instance, ΔNDCG=0.31 is achieved when applying *KStepM* and an additional normalisation step. The results demonstrate superior performance when compared to *LDA* with ΔNDCG=0.10 applied to complete set of resource instances. The main contribution consists of (i) a scalable method for efficiently generating structured dataset topic profiles combining and configuring suitable methods for NER, topic extraction and ranking as part of an experimentally optimised configuration, and (ii) the generation of structured dataset profiles for a majority of linked datasets according to established dataset description vocabularies. Dataset profiles generated through this approach can be explored in a stand-alone online explorer[7], top-k topic annotations are used as part of the LinkedUp dataset catalog [8], and more recently, topic profiles have been used to develop dataset recommendation techniques [12]. While it has been noted that meaningfulness and comparability of topic profiles can be increased when considering topics associated with certain resource types only, as part of additional work, resource type-specific dataset profiling approaches have been introduced [34].

2.3 Improving Entity Retrieval

While previous sections address the problem of discovering datasets, i.e. graphs representing potentially large amounts of entities, the entity-centric nature of the Web of data involves tasks related to entity and object retrieval [3, 35] or entity-driven text summarization [9]. Major search engine providers such as Google and Yahoo! already exploit entity-centric data to facilitate semantic search using knowledge graphs. In such scenarios, data is aggregated from a range of sources calling for efficient means to search and retrieve entities in large data graphs.

[5] http://vocab.deri.ie/void
[6] http://data.linkededucation.org/vol/
[7] http://data-observatory.org/lod-profiles/profile-explorer/

In particular, *entity retrieval* (also known as *Ad-Hoc Object retrieval*) [26, 35] aims at retrieving relevant entities given a particular entity-seeking query, resulting in a ranked list of entities [3]. By applying standard information retrieval algorithms, like BM2F, on constructed indexes over the textual descriptions (*literals*) of entities, previous works have demonstrated promising performance.

While there is a large amount of queries that are topic-based, e.g. '`U.S. Presidents`', rather than entity-centric, approaches like [35] have proposed retrieval techniques that make use of explicit links between entities for result or query expansion, for instance, *owl:sameAs* or *rdfs:seeAlso* statements. However, such statements are very sparse, particularly across distinct datasets.

[15] proposes a method for improving entity retrieval results by *expanding* and *re-ranking* the result set from a baseline retrieval model (BM25F). Link sparsity is addressed through clustering of entities (*x–means* and *spectral* clustering), based on their similarity, using both lexical and structural features. Thus, result sets retrieved through the baseline approach are expanded with related entities residing the same clusters as the result set entities. Subsequent re-ranking considers the similarity to the original query, and their relevance likelihood based on the corresponding entity type, building on the assumption that particulary query types are more likely result in certain result types (*query type affinity*). The clustering process is carried out as offline preprocessing, while the entity retrieval, expansion and re-ranking are performed online. An experimental evaluation on the BTC12 dataset [17], where the clustering process was carried out on a large set of entities (over 450 million), and using the SemSearch[8] query dataset shows that the proposed approach outperforms existing basslines significantly.

3 Crawling & Fusion of Entity-centric Web Markup

While the previous section has discussed approaches for exploiting entity-centric data from traditional knowledge graphs and linked data, here we turn towards structured Web markup as an emerging and unprecedented source of entity-centric Web data. Markup annotations embedded in HTML pages have become prevalent on the Web, building on standards such as RDFa[9], Microdata[10] and Microformats[11], and driven by initiatives such as schema.org, a joint effort led by Google, Yahoo!, Bing and Yandex.

The Web Data Commons [22], a recent initiative investigating a Web crawl of 2.01 billion HTML pages from over 15 million pay-level-domains (PLDs) found that 30% of all pages contain some form of embedded markup already, resulting in a corpus of 20.48 billion RDF quads[12]. The scale and upward trend of adoption[13] - the proportion of pages containing markup increased from 5.76%

[8] http://km.aifb.kit.edu/ws/semsearch10/
[9] RDFa W3C recommendation: http://www.w3.org/TR/xhtml-rdfa-primer/
[10] http://www.w3.org/TR/microdata
[11] http://microformats.org
[12] http://www.webdatacommons.org
[13] http://webdatacommons.org/structureddata/

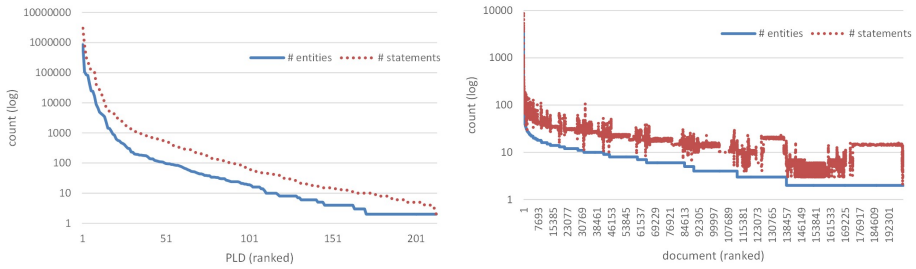
to 30% between 2010 and 2014 - suggest potential for a range of tasks, such as entity retrieval and knowledge base augmentation. However, facts extracted from embedded markup have different characteristics when compared to traditional knowledge graphs and linked data. In the following, we discuss first some case studies which investigate the coverage and distribution of Web markup for a particular set of entity types (Section 3.1), then we discuss apparent challenges (Section 3.2), and finally, we introduce current research which apply data fusion techniques to use Web markup data in the aforementioned tasks (Section 3.3).

3.1 Case Studies: Type-specific Coverage of Web Markup

As part of type-specific investigations [28, 29], we have investigated the scope, distribution and coverage of Web markup, specifically considering the cases of bibliographic data and of learning resource annotations. [28] provides a study of the adoption of Web markup for the annotation of bibliographic entities, being the first effort to investigate scholarly data extracted from embedded annotations. Utilising the WDC as largest crawl of embedded markup so far, the investigation considers all statements which describe entities (subjects) that are of type *s:ScholarlyArticle* or of any type but co-occurring on the same document with any *s:ScholarlyArticle* instance. Here and in the following we refer to the *http://schema.org* namespace as *s:*, and abbreviate *s:ScholarlyArticle* as *s:SchoArt*. Although there is a wide variety of types used for bibliographic and scholarly information, *s:SchoArt* is the only type which explicitly refers to scholarly bibliographic data. While this is a limitation with respect to recall, we followed this approach to enable a high precision of the analysed data within the scope of our study.

The extracted dataset contains 6,793,764 quads, 1,184,623 entities, 83 distinct classes, and 429 distinct predicates. Insights are provided with respect to frequent data providers, the adoption and usage of terms and the distribution across providers, domains and topics. The distribution of extracted data, spread across 214 distinct Pay-Level-Domains (PLDs), 38 Top-Level-Domains (TLDs) and 199,980 documents is represented in Figure 1. The blue (lower) line corresponds to the distribution of entities and the red and dashed (upper) line corresponds to the distribution of statements over PLDs/TLDs and documents. The number of entities/statements presented on the *y-axis* are plotted in the logarithmic scale. An apparent observation is the power law-like distribution, where usually a small amount of sources (PLDs, TLDs, documents) provide the majority of entities and statements. For example *springer.com* alone exposes a total of 850,697 entities and 3,011,702 statements. The same pattern can be identified for vocabulary terms, where few predicates are highly used, complemented by a long tail of predicates of limited use. With regard to the distribution across top-level-domains, a certain bias towards French data providers seems apparent based on some manual investigation of the top-k genres and publishers. Article titles, PLDs and publishers suggest a bias towards specific disciplines, namely Computer Science and the Life Sciences which mirrors a similar pattern in the

linked data world. However, the question to what extent this is due to the selective content of the Common Crawl or representative for schema.org adoption on the Web in general requires additional investigations.



(a) Distribution across PLDs (b) Distribution across HTML documents

Fig. 1. Distribution of entities/statements over PLDs and documents (from [28]).

[33] investigates the same corpus, yet towards the goal of understanding the adoption of *LRMI*[14] statements. The *Learning Resources Metadata Initiative (LRMI)* provides a *schema.org* extension tailored to the annotation of educational resources. In order to assess not only the coverage but also the evolution of LRMI statements, authors extracted subsets from the WDC2013 and WDC2014 datasets, by selecting all quads which co-occur with any of the LRMI vocabulary terms, such as *educationalAlignment*, *educationalUse*, *timeRequired*, or *typicalAgeRange*. The subsets under investigation contain 51,601,696 (WDC2013) respectively 50,901,532 (WDC2014) quads. The total number of entities in 2013 is 10,469,565 while in 2014 there are 11,861,807 entities, showing a significant growth in both cases. Regarding documents, we observe 3,060,024 documents in 2013 and 4,343,951 in 2014. Similarly to the case of bibliographic data, the distribution follows a power-law, where a small amount of providers (PLDs) provide large proportions of the data.

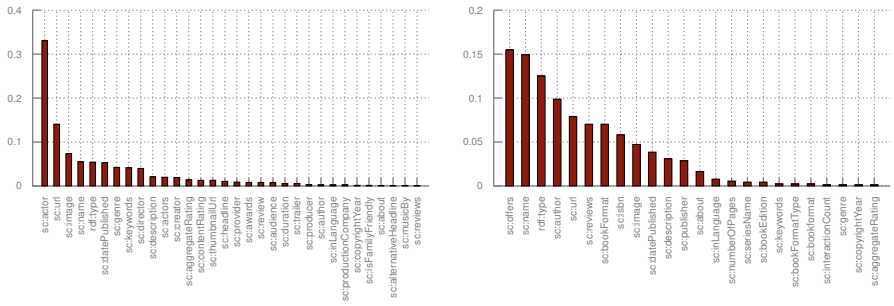
Findings from both studies suggest an uneven distribution of quads across documents and providers leading to potential bias in obtained entity-centric knowledge. On the other hand, the studies provide first evidence of a widespread adoption of even domain-specific types and terms, where in both cases, an inspection of the PLDs suggest that key data providers, such as publishers, libraries, or journals already embrace Web markup for improving search and interpretation of their Web pages. More exhaustive studies should consider, however, the use of focused crawls, which enable a more comprehensive study into the adoption of markup annotations in a respective domain.

[14] http://www.lrmi.net

3.2 Challenges

Initial investigations have shown the complementary nature of markup data, when compared to traditional knowledge bases, both at the entity level as well as the fact level, where the extent of additional information varies strongly between resource types. Though Web markup constitutes a rich and dynamic knowledge resource, the problem of answering entity-centric queries from entity descriptions extracted from embedded markup is a novel challenge, where the specific characteristics of such data pose different challenges [37] compared to traditional linked data:

- **Coreferences**: entities, particularly popular ones, are represented on a multitude of pages, resulting in vast amounts co-referring entity descriptions about the same entity. For instance, 797 entity descriptions can be obtained from WDC2014 which are of type *s:movie* and show a label (*s:name*) *Forrest Gump*.
- **Lack of explicit links**: RDF statements extracted from markup form a very sparsely linked graph, as opposed to the higher connectivity of traditional RDF datasets. This problem is elevated by the large amount of coreferences, where explicit links would facilitate the fusion of facts about the same entity from a variety of sources.
- **Redundant statements**: extracted RDF statements are highly redundant. For instance, Figure 2 presents a power law distribution for predicates observed from entity descriptions of type *s:Movie* and *s:Book*, where a few popular predicates occur in the vast majority of statements, followed by a long tail of infrequent predicates. Authors also observe that only a small proportion of facts are lexically distinct (60%), many of which are near-duplicates.
- **Errors**: as documented in [21], data extracted from markup contains a wide variety of syntactic and semantic errors, including typos or the misuse of vocabulary terms.



(a) s:Movie (b) s:Book

Fig. 2. Statement distribution across predicates for types *s:Movie* and *s:Book* (from [37]).

To this end, entity descriptions complement each other, yet sophisticated data fusion techniques are required in order to enable further exploitation of entity-centric knowledge from Web markup.

3.3 Exploiting Web Markup: Data Fusion and Knowledge Base Augmentation

Initial works such as the *Glimmer* search engine[15] have applied traditional *entity retrieval* techniques [4] to embedded markup (WDC). However, given the large amount of flat and highly redundant entity descriptions, practical use of search results obtained in that way is limited [37]. Key issues of such approaches include identity resolution as well as the vast amount of duplicates and near-duplicates. Therefore, the application of *data fusion* techniques is required to obtain a consolidated and correct entity description when answering entity-centric queries.

However, given the dynamics of Web markup data, the validity and correctness of a fact is usually of temporal nature [24]. For instance, the predicate *s:price* of a particular product offer is highly dynamic and its correctness depends strongly on the considered time frame. For these reasons, any data fusion approach would have to consider efficiency in order to enable frequent repetitions of the extraction pipeline, consisting of crawling, extraction and fusion. Considering the scale of large Web crawls such as the WDC, general data fusion strategies which are applied over the entire pool of data are impractical. This suggests a need for focused approaches, which are able to efficiently obtain fused entity descriptions for a given set of entity-seeking queries. For instance, for an entity-seeking query *'Iphone 6'* of type product, a *query-centric data fusion* approach will fuse all correct facts from an available corpus or crawl into a diverse entity description.

In [38], authors present *Clustering-Based Fact Selection(CBFS)* as an approach for query-centric data fusion of Web markup. Entity retrieval is conducted to provide a set of candidate facts for a given query. For this purpose, authors build a standard IR index of entity descriptions and apply the BM25 retrieval model on pseudo-key properties to obtain candidate entity descriptions. One major issue to address in the fact selection process, is the canonicalization of different surface forms, such as `Tom Hanks` and `T. Hanks`. To detect duplicates and near duplicates, authors cluster entity labels at the predicate level into n clusters $(c_1, c_2, \cdots, c_n) \in C$ using the X-Means algorithm [25]. Fact selection then considers a set of heuristics to enable the selection of correct and diverse facts from the candidate pool.

Experiments using the WDC2014 dataset indicate a comparably high precision 83.3% of this initial approach, showing a gain of 5.5% compared to a simple baseline. More recent work is concerned with building a supervised classification model for the data fusion step, based on a comprehensive feature set which considers relevance, quality and authority of sources, facts and entity descriptions.

[15] http://glimmer.research.yahoo.com/

To evaluate the potential of this approach for aiding knowledge base augmentation tasks, authors also measure the coverage gain by comparing obtained entity descriptions to their corresponding descriptions in DBpedia. It was found that 57% of the facts detected by CBFS do not exist in DBpedia with some of the facts corresponding to new predicates and some to already existing ones, which are not sufficiently populated. Considering only the predicates that exist in DBpedia, and the coverage gain is 33.4%. Currently ongoing research addresses the use of Web markup for tasks such knowledge base augmentation and temporal entity interlinking.

4 Conclusions

This paper provided an overview on selected works on retrieval, crawling and fusion of entity-centric Web data. While the heterogeneity and diversity of traditional linked data and knowledge graphs calls for efficient methods for dataset recommendation, profiling or entity retrieval (Section 2), we also investigated the exploitation of embedded Web markup data as emerging form of large-scale entity-centric data on the Web (Section 3). While an exhaustive literature review is out of scope of this paper, the focus here is on selected works covering a range of topics of relevance to general aim of retrieving entity-centric data from the Web.

Promising future directions are specifically concerned with the convergence of both sources of entity-centric knowledge discussed in this paper, for instance, by exploiting Web markup and data from Web tables for knowledge base augmentation. Interesting opportunities also emerge from the large-scale availability of markup and its use as unprecedented source of training data for supervised entitiy recognition, disambigutation or interlinking methods. The availability of explicit entity annotations at Web-scale enables the computation of a wide range of features which consider both, characteristics of unstructured Web documents as well as the embedded entity markup.

Acknowledgements. While all discussed works are joint research with numerous colleagues, friends and collaborators from a number of research institutions, the author would like to thank all involved researchers for the inspiring and productive work throughout the previous years. In addition, the author expresses his gratitude to all funding bodies that enabled the presented research through a variety of funding programs.

References

1. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. Dbpedia: A nucleus for a web of open data. In *International Semantic Web Conference (ISWC)*, pages 722–735, 2007.
2. C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.

3. R. Blanco, B. B. Cambazoglu, P. Mika, and N. Torzec. Entity recommendations in web search. In *Proceedings of the 12th ISWC*, pages 33–48, 2013.
4. R. Blanco, P. Mika, and S. Vigna. Effective and efficient entity search in rdf data. In *The Semantic Web–ISWC 2011*, pages 83–97. Springer, 2011.
5. K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD 2008, pages 1247–1250, New York, NY, USA, 2008. ACM.
6. S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7):107–117, 1998.
7. C. Buil-Aranda, A. Hogan, J. Umbrich, and P.-Y. Vandenbusshe. Sparql web-querying infrastructure: Ready for action? In *Proceedings of the 12th International Semantic Web Conference, Sydney, Australia*, 2013.
8. M. DAquin, A. Adamou, and S. Dietze. Assessing the educational linked data landscape. In *ACM Web Science 2013 (WebSci2013), Paris, France*. ACM, 2013.
9. G. Demartini, M. M. S. Missen, R. Blanco, and H. Zaragoza. Entity summarization of news articles. In *Proceeding of the 33rd ACM SIGIR*, pages 795–796, 2010.
10. S. Dietze, D. Taibi, and M. dAquin. Facilitating scientometrics in learning analytics and educational data mining the lak dataset. *Semantic Web Journal*, 2015.
11. S. Dietze, D. Taibi, H. Q. Yu, and N. Dovrolis. A linked dataset of medical educational resources. *British Journal of Educational Technology, BJET*, 46(5):1123–1129, 2015.
12. M. B. Ellefi, Z. Bellahsene, S. Dietze, and K. Todorov. Beyond established knowledge graphs-recommending web datasets for data linking. In A. Bozzon, P. Cudr-Mauroux, and C. Pautasso, editors, *ICWE*, volume 9671 of *Lecture Notes in Computer Science*, pages 262–279. Springer, 2016.
13. M. B. Ellefi, Z. Bellahsene, S. Dietze, and K. Todorov. Dataset recommendation for data linking: An intensional approach. In H. Sack, E. Blomqvist, M. d'Aquin, C. Ghidini, S. P. Ponzetto, and C. Lange, editors, *ESWC*, volume 9678 of *Lecture Notes in Computer Science*, pages 36–51. Springer, 2016.
14. B. Fetahu, S. Dietze, B. Pereira Nunes, M. Antonio Casanova, D. Taibi, and W. Nejdl. A scalable approach for efficiently generating structured dataset topic profi les. In F. Gandon and C. dAmato, editors, *In Proceedings of the 11th Extended Semantic Web Conference*. Springer, 2014.
15. B. Fetahu, U. Gadiraju, and S. Dietze. Improving entity retrieval on structured data. In *International Semantic Web Conference (1)*, volume 9366 of *Lecture Notes in Computer Science*, pages 474–491. Springer, 2015.
16. C. Guéret, P. T. Groth, C. Stadler, and J. Lehmann. Assessing linked data mappings using network measures. In *The Semantic Web: Research and Applications - 9th Extended Semantic Web Conference (ESWC)*, pages 87–102, 2012.
17. A. Harth. Billion Triples Challenge data set. Downloaded from http://km.aifb.kit.edu/projects/btc-2012/, 2012.
18. J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.
19. L. A. P. P. Leme, G. R. Lopes, B. P. Nunes, M. A. Casanova, and S. Dietze. Identifying candidate datasets for data interlinking. In *Proc. of the 13th ICWE*, pages 354–366, 2013.
20. G. Lopes, L. A. Paes Leme, B. Nunes, M. Casanova, and S. Dietze. Two approaches to the dataset interlinking recommendation problem. In *15th International Conference on Web Information System Engineering (WISE 2014)*, 2014.

21. R. Meusel and H. Paulheim. Heuristics for fixing common errors in deployed schema.org microdata. In *ESWC*, volume 9088 of *Lecture Notes in Computer Science*, pages 152–168. Springer, 2015.
22. R. Meusel, P. Petrovski, and C. Bizer. The webdatacommons microdata, rdfa and microformat dataset series. In *The Semantic Web–ISWC 2014*, pages 277–292. Springer, 2014.
23. B. P. Nunes, S. Dietze, M. A. Casanova, R. Kawase, B. Fetahu, and W. Nejdl. Combining a co-occurrence-based and a semantic measure for entity linking. In *10th Extended Semantic Web Conference (ESWC)*, pages 548–562, 2013.
24. Y. Oulabi, R. Meusel, and C. Bizer. Fusing time-dependent web table data. In *Proceedings of the 19th International Workshop on Web and Databases*, page 3. ACM, 2016.
25. D. Pelleg, A. W. Moore, et al. X-means: Extending k-means with efficient estimation of the number of clusters. In *ICML*, pages 727–734, 2000.
26. J. Pound, P. Mika, and H. Zaragoza. Ad-hoc object retrieval in the web of data. In *Proceedings of the 19th WWW*, pages 771–780, 2010.
27. G. Rabello Lopes, L. A. P. Paes Leme, B. Pereira Nunes, M. A. Casanova, and S. Dietze. Recommending tripleset interlinking through a social network approach. In *In the Proceedings of the 14th International Conference on Web Information System Engineering*, Lecture Notes in Computer Science. Springer, October 2013.
28. P. Sahoo, U. Gadiraju, R. Yu, S. Saha, and S. Dietze. Analysing structured scholarly data embedded in web pages. Apr. 2016.
29. P. Sahoo, U. Gadiraju, R. Yu, S. Saha, and S. Dietze. Analysing structured scholarly data embedded in web pages. In *Proceedings of the 25th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 2016.
30. M. Schmachtenberg, C. Bizer, and H. Paulheim. Adoption of the linked data best practices in different topical domains. In *Proc. of ISWC*, pages 245–260. 2014.
31. F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In C. L. Williamson, M. E. Zurko, P. F. Patel-Schneider, and P. J. Shenoy, editors, *WWW*, pages 697–706. ACM, 2007.
32. D. Taibi, S. Chawla, S. Dietze, I. Marenzi, and B. Fetahu. Exploring ted talks as linked data for education. *British Journal of Educational Technology*, (12283), 05/2015 2015.
33. D. Taibi and S. Dietze. Towards embedded markup of learning resources on the web: An initial quantitative analysis of lrmi terms usage. In J. Bourdeau, J. Hendler, R. Nkambou, I. Horrocks, and B. Y. Zhao, editors, *WWW (Companion Volume)*, pages 513–517. ACM, 2016.
34. D. Taibi, S. Dietze, B. Fetahu, and G. Fulantelli. Exploring type-specific topic profiles of datasets: a demo for educational linked data. In M. Horridge, M. Rospocher, and J. van Ossenbruggen, editors, *International Semantic Web Conference - Posters and Demos*, volume 1272 of *CEUR Workshop Proceedings*, pages 353–356. CEUR-WS.org, 2014.
35. A. Tonon, G. Demartini, and P. Cudré-Mauroux. Combining inverted indices and structured search for ad-hoc object retrieval. In *Proceedings of the 35th ACM SIGIR*, pages 125–134, 2012.
36. S. White and P. Smyth. Algorithms for estimating relative importance in networks. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 266–275, 2003.

37. R. Yu, B. Fetahu, U. Gadiraju, and S. Dietze. A survey on challenges in web markup data for entity retrieval. In *15th International Semantic Web Conference (ISWC2016)*, 2016.
38. R. Yu, U. Gadiraju, X. Zhu, B. Fetahu, and S. Dietze. Entity summarisation on structured web markup. In *The Semantic Web: ESWC 2016 Satellite Events*. Springer, 2016.
39. W. Yuan, E. Demidova, S. Dietze, and X. Zhou. Analyzing relative incompleteness of movie descriptions in the web of data: A case study. In M. Horridge, M. Rospocher, and J. van Ossenbruggen, editors, *International Semantic Web Conference - Posters and Demos*, volume 1272 of *CEUR Workshop Proceedings*, pages 197–200. CEUR-WS.org, 2014.