



Querying Heterogeneous Datasets on the Linked Data Web

Challenges, Approaches, and Trends

The growing number of datasets published on the Web as linked data brings both opportunities for high data availability and challenges inherent to querying data in a semantically heterogeneous and distributed environment. Approaches used for querying siloed databases fail at Web-scale because users don't have an a priori understanding of all the available datasets. This article investigates the main challenges in constructing a query and search solution for linked data and analyzes existing approaches and trends.

**André Freitas,
Edward Curry,
João Gabriel Oliveira,
and Seán O'Riain**
*Digital Enterprise Research
Institute*

The unprecedented availability of data promised by linked data¹ on the Web represents a major paradigm shift over the existing Web's structure. By building on Web infrastructure (URIs and HTTP), Semantic Web standards (such as the Resource Description Framework and RDF Schema [RDFS]), and vocabularies, linked data can effectively reduce barriers to data publication, consumption, and reuse, adding a rich layer of fine-grained, structured data to the Web. At its core, linked data exposes previously siloed databases as data graphs, which can be interlinked and integrated with other datasets, creating a global-scale interlinked dataspace.

However, linked data poses challenges inherent to querying highly

heterogeneous and distributed data. To query linked data on the Web today, users must first be aware of which exposed datasets potentially contain the data they want and what data model describes these datasets, before using this information to create structured queries. This query paradigm is deeply attached to the traditional perspective of structured queries over databases and doesn't suit the linked data Web's heterogeneity, distributiveness, or scale. It's impractical to expect Web data consumers to have a previous understanding of available linked datasets' structure and location. Letting users expressively query relationships in the data while abstracting them from the underlying data model is a fundamental problem

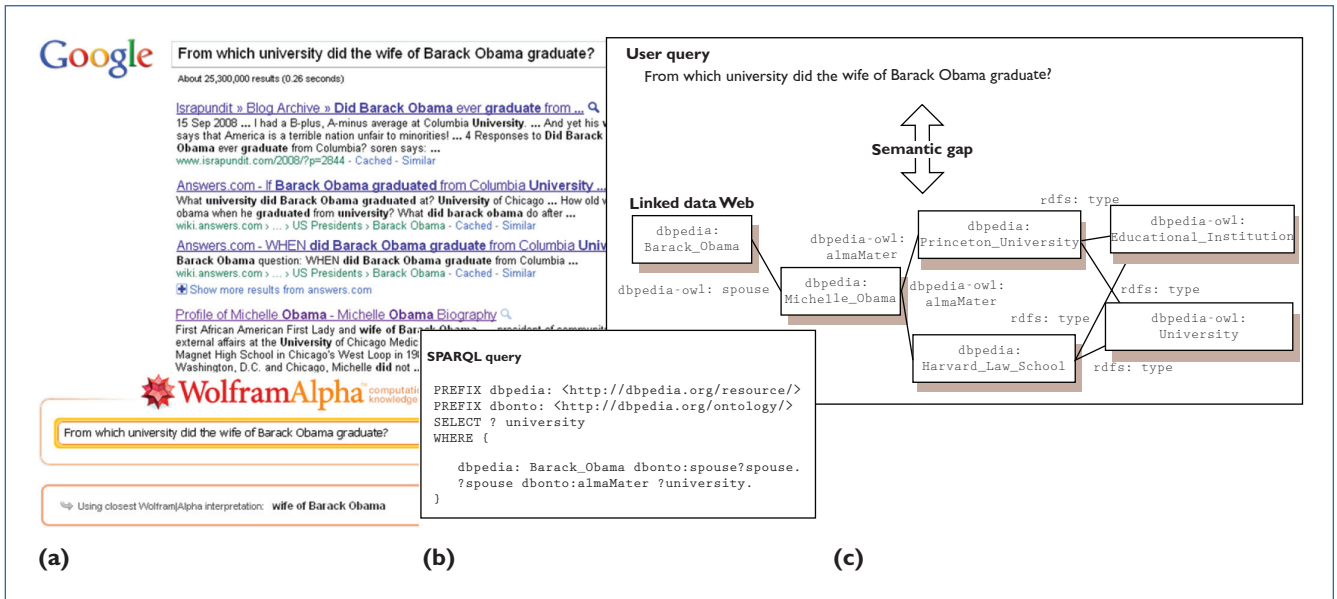


Figure 1. Querying data over the Web. We can see (a) a natural language query over two search engines; (b) the corresponding SPARQL representation; and (c) the semantic gap between the user’s information needs and the data representation.

for Web-scale data consumption, which, if not addressed, will ultimately limit linked data’s utility for consumers.

In addition to data model awareness, users querying linked data must master the syntax of structured query languages such as SPARQL. Most Web users aren’t comfortable with structured queries, thus creating a usability barrier for the linked data Web. From a user perspective, natural language queries emerge as a simple and intuitive alternative. Previous investigations have empirically confirmed natural language’s suitability for search and query tasks.²

This article provides a survey of existing approaches for searching and querying linked data on the Web, concentrating on how these approaches address the core challenges that emerge when heterogeneous datasets become exposed at Web-scale. Based on these challenges and approaches, this article also analyzes existing trends in the space. Linked data shares many of the objectives and challenges of data-spaces,³ a concept that expresses the recurring demand for dealing with heterogeneous, loosely connected, and distributed data sources. Data-spaces, however, don’t assume the support of linked data standards. Despite this difference, linked dataspaces and generic dataspaces share more commonalities than differences, and the analysis provided in this article can be transported to generic dataspaces.

Living in a Linked Dataspace

Linked data provides a data layer on the Web that represents objects and relations. The availability of Web-scale information in a structured and fine-grained representation could generate a paradigmatic shift in how applications and users consume data. Consider a journalist compiling a list of facts regarding public personalities and those personalities’ previous academic affiliations. The journalist can express his or her information needs as natural language queries, such as “From which university did the wife of Barack Obama graduate?” Document search engines can’t currently provide a level of query interpretation that could point directly to the final answer. With a traditional search engine, the journalist must navigate through the links and read the content of each candidate page the search engine returns. Modern search engines – such as Wolfram Alpha, which relies on manually curated structured knowledge sources – don’t provide a sufficiently comprehensive solution to answer this query (see Figure 1a).

The information that can answer this query is already available on the Web as linked data. However, to access it, users must know datasets’ location and structure, and the syntax of the SPARQL query language (see Figure 1b). Figure 1c shows the semantic gap between the user’s information needs expressed in a generic natural language query and the data representation

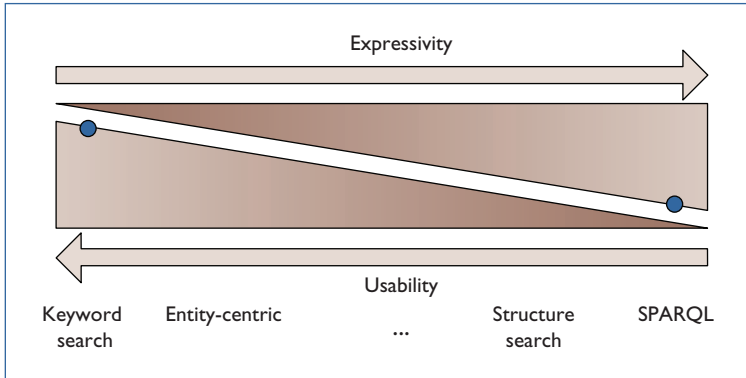


Figure 2. The expressivity–usability trade-off for querying over structured data. The blue dots indicate that an ideal query mechanism for linked data must provide both high expressivity and high usability. (This figure was adapted from previous work.²)

in the target dataset. The query’s terms and structure differ from the data representation in the dataset.

The linked data Web already contains valuable data in diverse areas, such as e-government, e-commerce, and the biosciences. Additionally, the number of available datasets has grown solidly since its inception.¹ The provision of intuitive and flexible query mechanisms that can approximate users from an unconstrained amount of data represents a fundamental challenge, which, if not addressed, could affect the linked data Web’s growth and adoption.

Challenges for Querying and Searching Linked Data

Search engines on today’s Web are based on variations of the *vector space model* (VSM). This model’s scalability and simplicity of use, based on keyword queries, defined its success as the de facto solution for search engines for the Web of documents. The VSM represents the contents of a collection of documents in a vector space built from terms present in the collection. Traditional VSM solutions lack the representation of structure information needed for data queries. This is reflected in their alternative name, “bag-of-words” approaches.

In the (semi-)structured data world, the relationships between entities in a dataset are fundamental to the model the dataset represents. Today, structured query languages are the standard way to query structured data. Structured queries are essentially built from two components: the query language’s *syntax* and the *elements*

(*entities* and *relationships*) of the data model behind the dataset. The structured query approach fails on the linked data Web, however, because the Web’s scale makes it infeasible for users to become aware of the structure of datasets to query them.

Consequently, the linked data Web demands approaches that can combine VSMS’ usability and scalability with the expressivity required to query (semi-)structured data, bridging the semantic gap between users and the linked data Web. Figure 2 depicts the trade-off between query expressivity and usability; existing approaches are positioned along an expressivity–usability spectrum. This trade-off is a consequence of the semantic gap for linked data queries. Ideally, a query mechanism for linked data must provide both high expressivity and high usability (the blue dots in the figure). It should also employ a level of semantic interpretation and matching not present in standard search and query approaches.

Previous works have proposed various solutions to address these challenges. To understand their strengths and limitations, we present five core challenge dimensions:

- *Query expressivity* is the ability to query datasets by referencing elements in the data model structure, as well as to operate over the data (aggregate results, express conditional statements, and so on).
- *Usability* allows for an easy-to-operate, intuitive, and task-efficient query interface.
- *Vocabulary-level semantic matching* is the ability to semantically match user query terms to dataset vocabulary-level terms.
- *Entity reconciliation* matches entities expressed in the query to semantically equivalent dataset entities.
- *Semantic tractability mechanisms* improve on the ability to answer queries not supported by explicit dataset statements (for example, “Is Natalie Portman an Actress?” can be supported by the statement “Natalie Portman starred Star Wars,” instead of an explicit statement “Natalie Portman occupation Actress,” which might not be present in the dataset).

These challenges concentrate on the core usability and semantic aspects necessary to address the usability–expressivity trade-off.

Existing Approaches

Three high-level categories of approaches for querying linked data exist: approaches employing strategies inherited from the information retrieval (IR) space in which keyword search is mixed with elements from structure queries; approaches focusing on natural language queries; and structured SPARQL queries over distributed datasets. Here, we focus on the usability and semantic matching problems, thus analyzing approaches from the first two categories.

Information Retrieval Approaches

We can categorize IR approaches according to index type, which includes *entity-centric search* approaches and *structure search* approaches. Although both types provide hybrid search interfaces that merge keyword search with dataset structure elements, only structure search targets indexing strategies focusing on addressing the expressivity-usability trade-off at the index construction level.

Entity-centric search. Entity-centric approaches let users search for entities (*instances* and *classes*) in datasets, employing VSM variations to index those entities. Existing approaches range from less expressive queries, based on keyword search over textual information associated with the dataset entities, to *star-shaped queries* and *hybrid queries* (that is, queries mixing keyword search, and structured queries centered on an entity).

The Semantic Web Search Engine (SWSE) is a search and query service that implements an architecture with components for crawling, integrating, indexing, querying, and navigating over multiple data sources.⁴ The system architecture's main components include query processing, ranking, an index manager, and an internal data store (YARS2), which focuses on scalability issues to enable federated queries over linked data. SWSE uses an approach called ReConRank to rank entities;⁴ this approach adapts the Page-Rank algorithm to work over RDF datasets, propagating dataset-level scores – computed from interlinking patterns – to data-level entities. The Scalable Authoritative OWL Reasoner (SAOR) provides an RDFS and partial Web Ontology Language (OWL) reasoning engine to address scalability issues.⁴ SAOR applies reasoning only on dataset fragments supported by an authoritative ontological definition.

Sindice is a search and query service for the linked data Web that ranks entities according to the incidence of keywords associated with them.⁵ It uses a node-labeled tree model to represent the relationship between datasets, entities, attributes, and values. Similarly to SWSE, Sindice provides a comprehensive entity-centric search and indexing approach. Figure 3a depicts Sindice's architecture.

Entity-centric search approaches have developed comprehensive data management strategies for linked data on the Web, providing the infrastructure for managing the complete crawl-index-search cycle. These approaches also developed services complementary to the entity-centric search process that let users either visually explore (via Visinav⁴ and Sigma⁵) or execute full structured SPARQL queries over the crawled data. Entity-centric approaches avoid major changes

Entity-centric search approaches have developed comprehensive data management strategies for linked data on the Web.

in standard indexing strategies, inheriting index and search optimization mechanisms present in existing VSM frameworks. These approaches have avoided tackling the expressivity-usability trade-off by aggregating multiple query interfaces; in practice, to execute expressive queries, users must be aware of the vocabularies behind the datasets. In addition, most entity-centric approaches have only limited evaluation in terms of search result quality.

Structure search. Structure search engines improve keyword queries' expressivity, extending existing inverted list indexes to represent structure information present in datasets. The main difference between entity-centric search and structure search is that the latter improves query expressivity with support from the extended index.

The search engine Semplore uses a hybrid query formalism that combines keyword search with structured queries (that is, a subset of SPARQL).⁶ Semplore uses position-based indexing

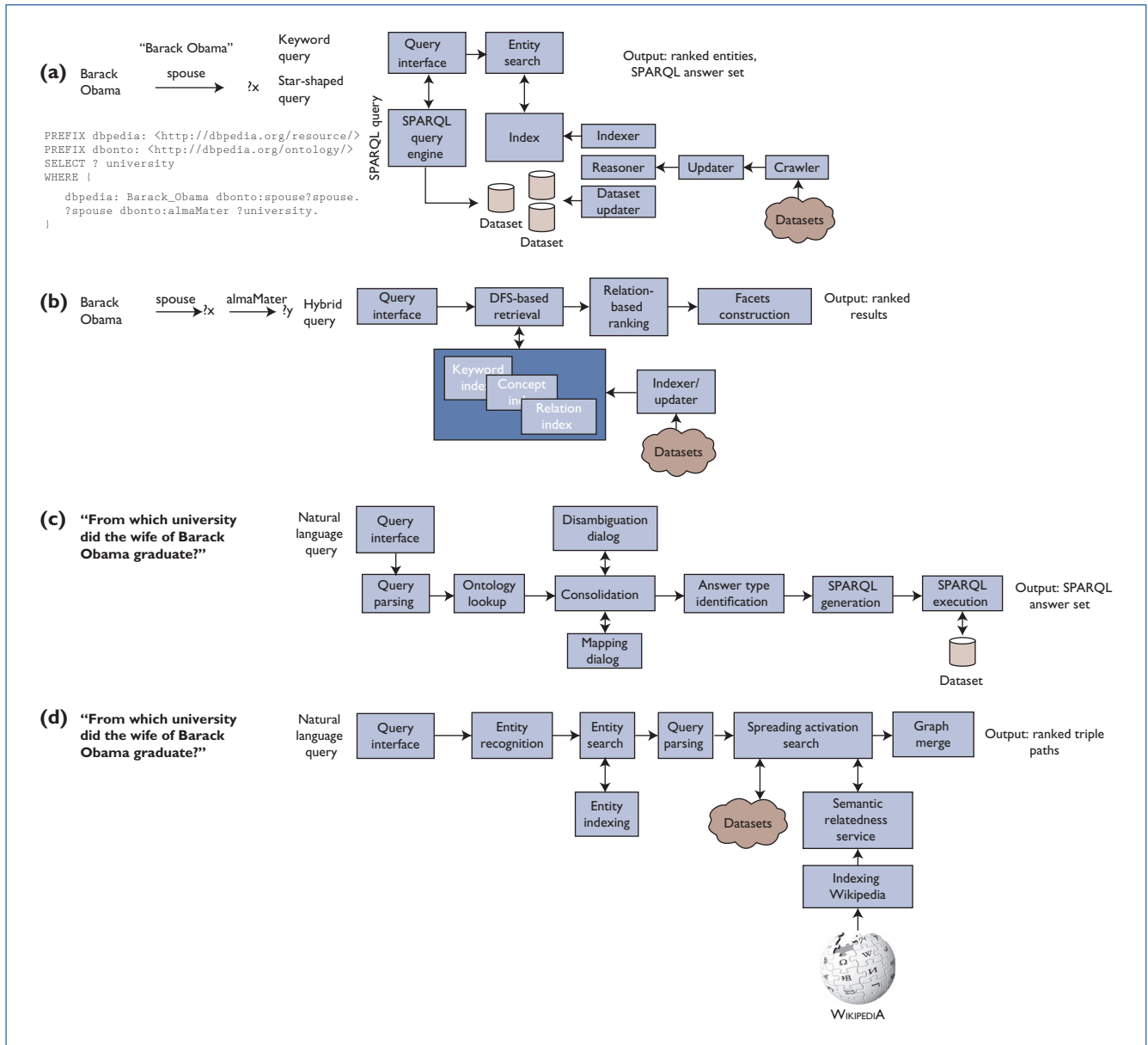


Figure 3. Examples of linked data search/query systems. We can see the high-level architecture components for (a) Sindice (entity-centric search), (b) Semplore (structure search), (c) FREyA (question answering), and (d) Treo (best-effort natural language).

to index relations and join triples. It relies on three types of inverted indexes: *keyword*, *concept*, and *relation*. Semplore also explores user feedback strategies for improving search, providing a faceted and navigational interface. Figure 3b depicts Semplore’s high-level architecture. Xin Dong and Alon Halevy propose an approach for indexing triples to enable queries that combine keywords and dataset structure elements.⁷ To provide a more flexible semantic matching, the authors propose four structured index types based on the introduction of additional structure information and semantic enrichment in

the inverted lists. Taxonomies associated with the dataset vocabularies are used as a semantic enrichment strategy.

Structure search approaches target the expressivity-usability trade-off by modifying and extending traditional inverted index structures. They introduce a limited level of semantic matching by taking into account the terminology-level information present in datasets or by enriching the index with related terms using WordNet. No comprehensive evaluation of the search results’ quality exists, making it unclear how these approaches

perform in addressing the expressivity–usability trade-off.

Natural Language Approaches

Approaches in the literature based on natural language queries target query mechanisms with high usability and expressivity. Although some approaches focus on the *question-answering* (QA) problem, in which, similarly to databases, precise answers are expected as the output, others focus on a *best-effort* scenario that returns a ranked list of results.

Question answering. The investigation of QA systems focuses on the problem of allowing users to query data using natural language queries. As opposed to IR techniques’ best-effort nature, QA systems target crisp answers, as with structured queries over databases. Work on QA approaches investigates the interpretation of users’ information needs expressed as natural language queries, applying natural language processing (NLP) techniques to parse queries and match them with dataset structures. Substantial research efforts have focused on this problem. We look at two recent works on open domain linked data.

PowerAqua is a QA system that uses PowerMap, a hybrid matching algorithm comprising terminology-level and structural schema-matching techniques with the assistance of large-scale ontological or lexical resources.⁸ In addition to the ontology structure, PowerMap uses WordNet-based similarity approaches as a semantic approximation strategy. Exploring user interaction techniques, FREyA is a QA system that employs feedback and clarification dialogs to resolve ambiguities and improve the domain lexicon with users’ help.⁹ Compared to PowerAqua, FREyA delegates a large part of the semantic matching and disambiguation process to users. User feedback enriches the semantic matching process by allowing manual entries of query-vocabulary mappings. Figure 3c depicts FREyA’s high-level architecture.

Compared to IR-based approaches, QA approaches aim toward more sophisticated semantic matching techniques because they target queries with high expressivity and don’t assume users are aware of the dataset representations (high usability). In contrast to entity-centric and structure search approaches, QA systems have a strong tradition of evaluating

results’ quality, having concentrated less on performance and scalability issues. Traditionally, QA approaches have focused on limited semantic matching (WordNet-based) strategies, making them unable to cope with the Web environment’s heterogeneity. Most QA approaches apply limited semantic matching techniques (for example, synonymic, taxonomic similarity) for matching query terms to dataset terms. In addition, they depend on resources that are manually created (WordNet) and difficult to expand across different domains.

Best-effort natural language interfaces. Some recent approaches aim to merge natural language queries’ expressivity and usability with IR models’ scalability and best-effort nature, targeting a best-effort natural language search mechanism. As in QA systems, users can still enter full natural language queries; however, instead of targeting crisp answers, these approaches return an approximate ranked list of results.

Treo is a natural language query mechanism for linked data that uses semantic relatedness measures derived from Wikipedia to match query terms to dataset terms.¹⁰ The use of semantic relatedness measures allows the quantification of the *semantic proximity* between two terms, using semantic information which is embedded in large textual resources available on the Web such as Wikipedia. Wikipedia-based semantic relatedness measures address previous limitations of WordNet-based semantic matching. Treo’s approach combines entity search, spreading activation search, and semantic relatedness to navigate over the linked data Web graph, semantically matching the parsed user query to the data representation in the datasets. Figure 3d depicts Treo’s components.

In prior work, we generalized the principles of the Treo approach by constructing a distributional semantic space (T-Space) for linked datasets.¹¹ We built this space using a distributional semantic model based on statistical semantic information derived from Wikipedia. This model enables flexible semantic matching in the search process (we discuss distributional semantic models in more detail later). The definition of the T-Space provides a principled representation of datasets focused on addressing the expressivity–usability trade-off.

Table 1. Strategies employed by each approach to address existing linked data querying challenges.

Approaches		Challenges				
		Usability	Query expressivity	Vocabulary-level semantic matching	Entity reconciliation	Improvement of semantic tractability
Information retrieval	Entity-centric (SWSE/Visinav, ⁴ Sindice/Sigma ⁵)	High	Keywords, star-shaped, ⁵ SPARQL	No	OWL:same as, OWL: Inverse Functional	Contextual ⁵ and best-effort authoritative reasoning ⁴ (RDFS and OWL subset)
	Structure indexes (Semplore, ⁶ Dong and Halevy ⁷)	Medium	Keywords, conjunctive/path queries	Taxonomy indexing, descriptions, and associations enrichment; ⁶ WordNet synonym ⁷	No	No
Natural language approaches	Question-answering systems (PowerAqua, ⁸ Freya ⁹)	High	Natural language queries	WordNet, ontology structure, ⁸ user enrichment ⁹	Dataset look-up, user feedback	WordNet-based semantic similarity
	Natural language search (Treo, ¹⁰ Treo T-Space) ¹¹	High	Natural language queries (no operators)	Wikipedia-based, semantic relatedness, Wikipedia Link Measure (WLM), ¹⁰ Explicit Semantic Analysis (ESA) ¹¹	TF/IDF (instances) and ESA (classes) ¹¹	Wikipedia-based, semantic relatedness, WLM, ¹⁰ ESA ¹¹
Structured queries	SPARQL	Low	High	No	No	No

*Cell shading reflects the level at which the proposed strategies address the challenges (light shading represents less coverage; dark shading represents greater coverage).

Best-effort natural language search approaches provide a more robust semantic matching approach. However, they relax expectations in terms of query results, delegating the results' final assessment to end users. Similarly to QA systems, these approaches have concentrated on evaluating search results' quality. Table 1 lists how each category addresses key usability and semantic matching challenges. It also summarizes existing approaches' strengths and limitations, depicting their complementary aspects. Finally, it analyzes how key features in existing systems can align to provide a comprehensive linked data query solution.

Taming Data Heterogeneity

Our analysis of the existing approaches and how they address the challenges of querying linked data over the Web defines a landscape for the key features likely present in search and query mechanisms over linked dataspace. Seven key search and query features emerge from this analysis as clear trends. Table 2 summarizes each feature's impact in the various challenge dimensions. We grouped the features by three main architectural elements: *user interaction and interface, query processing and search, and index.*

Table 2. Key search and query features and their impact on the set of challenges.

Architectural elements	Key features	Challenges				
		Usability	Query expressivity	Vocabulary-level semantic matching	Entity reconciliation	Improvement of semantic tractability
User interaction and interface	Complementary search and query services	High	High	—	—	—
	User interaction and feedback mechanisms	—	—	Medium	Medium	Medium
Query processing and search	Best-effort query model	High	High	Medium	Medium	—
	Use of natural language processing techniques	High	High	Medium	—	—
Index	Distributional semantic model	High	High	High	Medium	High
	Use of external knowledge sources for semantic enrichment	Medium	—	High	High	High
	Integrated entity reconciliation techniques	—	—	—	High	—

Complementary Search and Query Services

Entity-centric search, keyword-based search, natural language queries, and structured SPARQL queries represent complementary search and query services that might suit users in different tasks and purposes. Search and query platforms should explore this complementary aspect with regard to heterogeneous data to enable users to switch among different search and query strategies. SWSE and Sindice are exploring this trend; however, the availability of natural language queries is a key feature not present in these systems. As part of the search and query features, users should be able to explore, understand, and refine search results by relying on navigational, browsing, and filtering capabilities integrated into the process (this functionality is present in SWSE, Sindice, and Semplore).

User Interaction and Feedback Mechanisms

The presence of ambiguity and incomplete information is intrinsic to the search and query process. As already explored in systems such as

FREyA and Semplore, user feedback can help resolve ambiguities, enrich an application’s semantic model, and filter and post-process results.

Best-Effort Query Model

In “If You Have Too Much Data, then ‘Good Enough’ Is Good Enough,”¹² Pat Helland summarizes the mindset shift that must occur in heterogeneous and distributed data environments, where many still expect the accurate and crisp results common for siloed databases. The challenge of building query solutions with high usability and expressivity is coping with the data’s semantic heterogeneity at Web-scale; this demands relaxing our expectations of the results into a best-effort solution. Ranked lists of results in which users can assess those results’ suitability are widely used in document search engines; Web users have been extensively exposed to this approach and are thus familiar with best-effort search models. However, although document search engines can potentially return a long list of candidate documents, best-effort query

mechanisms for linked data should leverage the structure and types present in the data to target more concise answer sets.

Note also the need to provide a supporting context around the answers that can help users assess the data's correctness. In the Treo approach, the path in the dataset generated during the querying process provides contextual information for users. A best-effort approach can live together with database operations, such as aggregations, via data filtering mechanisms that let users remove incorrect entries from the results (for example, using the associated type information).

Natural Language Processing Techniques

For many years, the difficulties associated with the hard constraints of the QA problem have overshadowed the potential for applying NLP techniques for queries. NLP has developed a large set of techniques and tools for parsing and analyzing users' information needs expressed as natural language queries. Different flavors of syntactic parsers, morphological analyzers, and named entity recognition techniques are widely and effectively employed in different problems and in QA systems and natural language search interfaces (for example, PowerAqua, FREyA, Treo, and Treo T-Space). Recently, NLP techniques' efficacy was demonstrated in the IBM Watson system,¹³ which outperformed its human contestant in a "Jeopardy" challenge. Watson heavily leverages standard NLP techniques to build a complex information extraction and search pipeline. Search and query mechanisms can explore NLP techniques to provide expressive and intuitive query interfaces.

Distributional Semantic Model

The difficulty in effectively providing a robust semantic matching solution has been associated with a level of semantic interpretation that depends on fundamental and hard problems in artificial intelligence, such as commonsense knowledge representation and reasoning. Recently, however, distributional semantic approaches are emerging as grassroots solutions to provide robust semantic matching by leveraging the use of semantic information embedded in large amounts of Web corpora.

Distributional semantic models assume that the context surrounding a given word in a text provides important information about its meaning.¹⁴ Distributional semantics focuses

on constructing a semantic representation of a word based on the statistical distribution of word co-occurrence in texts. The availability of high-volume and comprehensive Web corpora has made distributional semantic models a promising approach for building and representing meaning. However, the simplification of distributional semantic models implies some constraints on its use as a semantic representation. Distributional semantic models are suitable for computing semantic relatedness, which can act as a best-effort solution for providing robust semantic matching solutions for linked data queries (present in the Treo T-Space system).

External Knowledge Sources for Semantic Enrichment


The availability of large amounts of unstructured text and structured data on the Web can help to bootstrap a level of semantic interpretation based on available open and domain-specific knowledge. It is possible to address the volume of unstructured text corpora necessary to build distributional semantic models by using comprehensive knowledge sources available on the Web, such as Wikipedia (present in the Treo and Treo T-Space systems). In addition, it is possible to use the semantically rich entity structure of data sources such as DBpedia (<http://dbpedia.org>), YAGO (www.mpi-inf.mpg.de/yago-naga/yago/), and Freebase (www.freebase.com) as a general-purpose entity and entity typing system that can easily integrate to the target datasets to provide a minimum level of structured commonsense knowledge, and which can later be used to improve semantic interpretation and tractability. RDF's standardized graph-based format facilitates the reuse and integration of existing data sources into target datasets.

Integrated Entity Reconciliation Techniques

Existing search and query approaches haven't fully integrated current solutions (for example, similarity-based) for entity reconciliation (ER) into the index construction process, leaving a functional gap that must be addressed by future query mechanisms by applying more principled ER solutions.

The emergence of heterogeneous and distributed Web-scale data environments, in contrast to small, controlled schema databases, fundamentally shifts how users query data. Our analysis of the state of the art shows that existing

approaches based on IR and natural language query interfaces have complementary features, which, if combined, can provide solutions to existing usability and semantic matching challenges. Some of these features suggest important trends that will become key functionalities in future search and query mechanisms.

The challenges involved in constructing effective query mechanisms for Web-scale data offer an opportunity to converge three very active research areas, bringing together databases, IR, and natural language processing. The results emerging from this convergence will profoundly affect how humans interact with information. 

Acknowledgments

This work has been funded by Science Foundation Ireland under grant number SFI/08/CE/I1380 (Lion-2). We thank the reviewers and editors for their careful and valuable feedback.

References

1. T. Berners-Lee, "Linked Data Design Issues," 2009; www.w3.org/DesignIssues/LinkedData.html.
2. E. Kaufmann and A. Bernstein, "Evaluating the Usability of Natural Language Query Languages and Interfaces to Semantic Web Knowledge Bases," *J. Web Semantics: Science, Services, and Agents on the World Wide Web*, vol. 8, 2010, pp. 377–393.
3. M. Franklin, A. Halevy, and D. Maier, "From Databases to Dataspaces: A New Abstraction for Information Management," *SIGMOD Record*, vol. 34, no. 4, 2005, pp. 27–33.
4. A. Hogan et al., "Searching and Browsing Linked Data with SWSE: The Semantic Web Search Engine," *J. Web Semantics*, to appear, 2011.
5. R. Delbru, S. Campinas, and G. Tummarello, "Searching Web Data: An Entity Retrieval and High-Performance Indexing Model," *J. Web Semantics*, to appear, 2011.
6. H. Wang et al., "Semplore: A Scalable IR Approach to Search the Web of Data," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 7, no. 3, 2009, pp. 177–188.
7. X. Dong and A. Halevy, "Indexing Dataspaces," *Proc. 2007 ACM SIGMOD Int'l Conf. Management of Data*, ACM Press, 2007, pp. 43–54.
8. V. Lopez, E. Motta, and V. Uren, "PowerAqua: Fishing the Semantic Web," *Proc. 3rd European Semantic Web Conf. (ESWC 04)*, vol. 4011, Springer, 2004, pp. 393–410.
9. D. Damjanovic, M. Agatonovic, and H. Cunningham, "FREyA: An Interactive Way of Querying Linked Data Using Natural Language," *Proc. 1st Workshop on Question Answering over Linked Data (QALD-1), Collocated with the 8th Extended Semantic Web Conf. (ESWC 11)*, 2011.
10. A. Freitas et al., "Querying Linked Data Using Semantic Relatedness: A Vocabulary Independent Approach," *Proc. 16th Int'l Conf. Applications of Natural Language to Information Systems (NLDB 11)*, Springer, 2011, pp. 40–51.
11. A. Freitas et al., "A Multidimensional Semantic Space for Data Model Independent Queries over RDF Data," *Proc. 5th IEEE Int'l Conf. Semantic Computing (ICSC 11)*, IEEE Press, 2011, pp. 344–351.
12. P. Helland, "If You Have Too Much Data, then 'Good Enough' is Good Enough," *Comm. ACM*, vol. 54, no. 6, 2011.
13. D. Ferrucci et al., "Building Watson: An Overview of the DeepQA Project," *AI Magazine*, vol. 31, no. 3, 2010, pp. 59–79.
14. P.D. Turney and P. Pantel, "From Frequency to Meaning: Vector Space Models of Semantics," *J. Artificial Intelligence Research*, vol. 37, 2010, pp. 141–188.

Andre Freitas is a PhD student at the Digital Enterprise Research Institute (DERI), National University of Ireland, Galway. His main research interests include semantic search, linked data queries, and provenance. Freitas has a BSc in computer science from the Federal University of Rio de Janeiro (UFRJ). Contact him at andre.freitas@deri.org.

Edward Curry is a research leader at the Digital Enterprise Research Institute (DERI), National University of Ireland, Galway, and an adjunct lecturer at NUI, Galway. His projects include studies of enterprise linked data, energy informatics, semantic information management, and community-based data curation. Curry has a PhD from NUI, Galway. Contact him at ed.curry@deri.org.

João Gabriel Oliveira is a research intern at the Digital Enterprise Research Institute (DERI), National University of Ireland, Galway. His main research interests include natural language processing and semantic search. Oliveira is finishing a BSc in computer science at the Federal University of Rio de Janeiro (UFRJ). Contact him at joao.deoliveira@deri.org.

Sean O'Riain leads the e-business domain at the Digital Enterprise Research Institute (DERI), National University of Ireland, Galway. His research interests include the application of natural language processing and Semantic Web technologies and standards in business information systems. O'Riain has an MSc in distributed information retrieval from the National University of Ireland, Galway. Contact him at sean.oriain@deri.org.