Why the Data Train Needs Semantic Rails

Krzysztof Janowicz
University of California, Santa Barbara, USA
janowicz@ucsb.edu

Frank van Harmelen
Vrije Universiteit Amsterdam, The Netherlands
frank.van.harmelen@cs.vu.nl

James A. Hendler
Rensselaer Polytechnic Institute, Troy, NY, USA
hendler@cs.rpi.edu

Pascal Hitzler
Wright State University, Dayton,Ohio, USA
pascal.hitzler@wright.edu

Abstract

While catchphrases such as big data, smart data, data-intensive science, or smart dust highlight different aspects, they share a common theme: Namely, a shift towards a data-centric perspective in which the synthesis and analysis of data at an ever-increasing spatial, temporal, and thematic resolution promises new insights, while, at the same time, reducing the need for strong domain theories as starting points. In terms of the envisioned methodologies, those catchphrases tend to emphasize the role of predictive analytics, i.e., statistical techniques including data mining and machine learning, as well as supercomputing. Interestingly, however, while this perspective takes the availability of data as a given, it does not answer the question how one would *discover* the required data in today's chaotic information universe, how one would *understand* which datasets can be meaningfully *integrated*, and how to *communicate* the results to humans and machines alike. The Semantic Web addresses these questions. In the following, we argue why the data train needs semantic rails. We point out that making sense of data and gaining new insights works best if inductive and deductive techniques go hand-in-hand instead of competing over the prerogative of interpretation.

Introduction and Motivation

Typically, the data universe and its impact on science and society are discussed from an analysis perspective, i.e., how to transform data into insights instead of drowning in information. Equally important, however, are questions of how to effectively publish data and break up data silos, how to retrieve data, how to enable the exploration of unfamiliar datasets from different domains, how to access provenance information, how to determine whether datasets can be meaningfully reused and integrated, how to prevent data from being misunderstood, how to combine data with processing services and workflows on-the-fly, and finally how to make them readable and understandable by machines and humans.

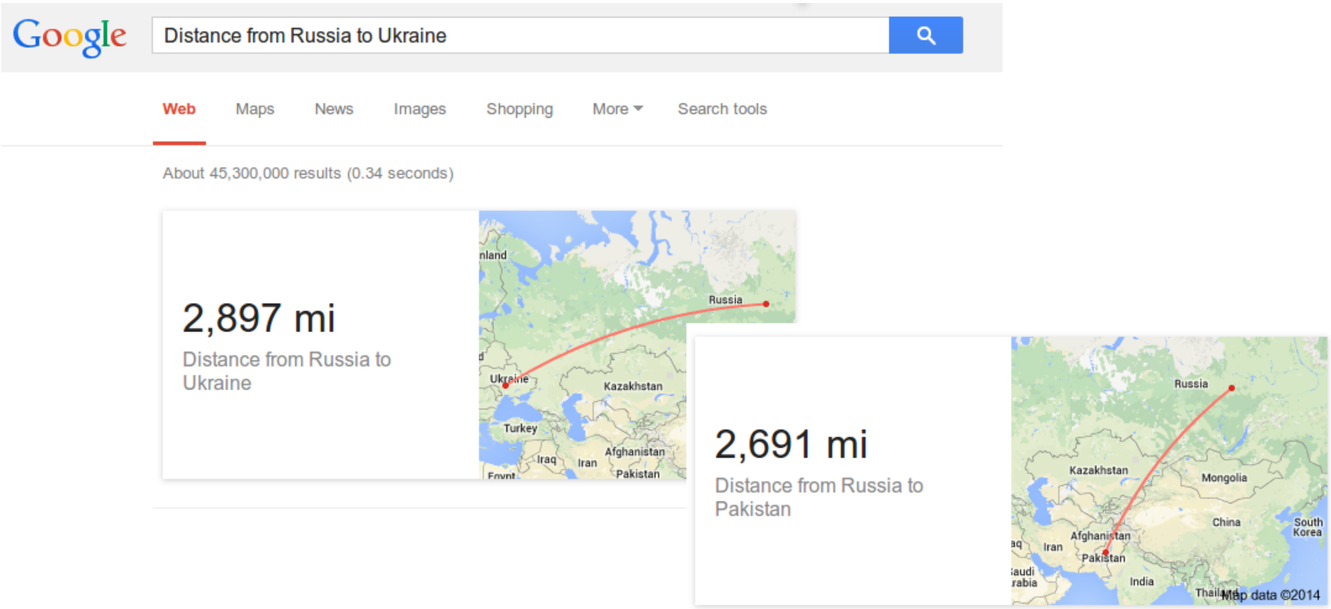
In other words, turning data into insights requires an infrastructure for publishing, storing, retrieving, reusing, integrating, and analyzing data. In the following, we will argue that the Semantic Web provides such an infrastructure and is entirely based on open and well-established standards. Together with the paradigm of making data *smart*, these standards ensure the longevity, independence, and robustness of this infrastructure. Instead of presenting an all-encompassing survey, we will focus on selected aspects that are of particular interest to data-intensive science (Hey et al. 2009) thereby illustrating the value proposition of Semantic Technologies and ontologies. While we will discuss use cases from the life sciences, natural sciences, and social sciences, the following humorous example shall act as an illustrative starting point.

In a recent guest article for the Washington Post, researchers discussed a relation between the believed geographic distance between the USA and Ukraine and the willingness to intervene in the 2014 Crimea conflict. Participants were more likely to support an US intervention the less accurately they could position Ukraine on a map. The authors state that 'the further our respondents thought that Ukraine was from its actual location, the more they wanted the U.S. to intervene militarily'(Dropp, Kertzer, and Zeitzoff 2014).

Let us assume that a researcher would like to investigate whether one could generalize this finding by testing a related hypothesis, namely that *citizens of a given country are more likely to support their government to intervene in another country depending on the (increasing) distance between both countries*. Naively, one could collect data about the distances between countries such as Russia, Ukraine, Armenia, Pakistan, and so forth, and then interview citizens from those countries to test the hypothesis. In fact, such distance data is easily available and a distances matrix between the roughly 200 countries can be create for further analysis in any statistical software environment.

Whatever the resulting correlation and significance level may be, the results will neither be meaningful nor reproducible. First, most datasets (including search and question answering engines such as Google or Wolfram Alpha) will compute the distance based on the country centroids, thereby producing misleading or even meaningless results. For instance, Russia will appear closer to Pakistan than to the Ukraine; see figure 1. Second, without a formal definition of *country*, it will be difficult to reproduce the experiment. Was the collected data about states, countries, or nations? Did the dataset contain the 193 United Nations member states, or other states as well. If the used dataset was created in

Final draft (May 2014).



Distance from Russia to Ukraine

About 45,300,000 results (0.34 seconds)

2,897 mi

Distance from Russia to Ukraine

2,691 mi

Distance from Russia to Pakistan

Figure 1: Distances between countries according to Google search.

Pakistan, Armenia may be missing as Pakistan does not recognize Armenia as a state. Similarly, the centroids will also vary depending on the origin of the dataset. For instance, data from India, China, and Pakistan will show different centroids as each of them claims certain parts of the Kashmir region as their territory (The Economist 2012). As a consequence, Google Maps, for instance, draws different borders depending on whether a user is accessing the service from India or the US.

To avoid such (and many other) difficulties, it is crucial to provide humans and machines with additional information. For instance, the fact that the used coordinates represent points (centroids) and not polygons, that topological information about neighboring states is available (Russia and Ukraine share a border, while Russia and Pakistan do not), that 17% of Ukraine's population is of Russian ethnicity, that the UN list of member states has been used as extensional definition of the term *country*, that the centroids were recorded by a mapping agency in India, and so forth. Fortunately, such information is available – namely as semantically-enabled Linked Data.

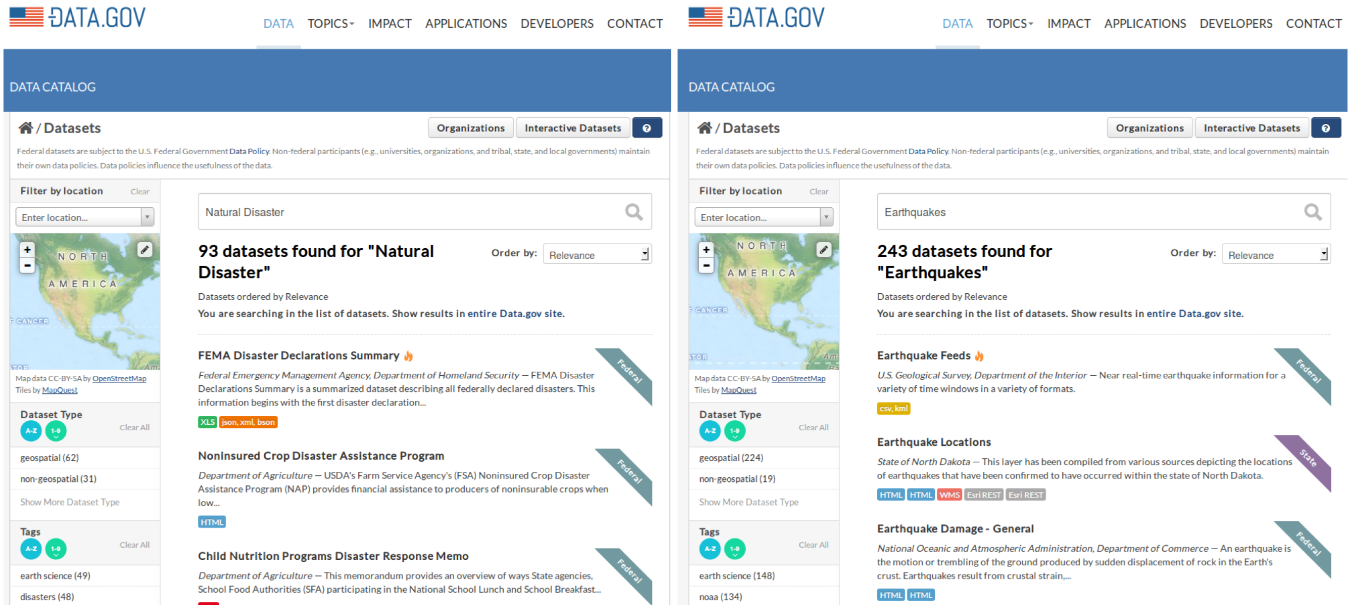
One could now argue that domain experts would be aware of the discussed difficulties and would therefore not use centroid data, know about ongoing territorial disputes, and so forth. In the big data age, however, *synthesis is the new analysis*. New insights are gained from integrating and mining multi-thematic and multi-perspective data from highly heterogeneous resources across domains and disciplines.

Synthesis is the New Analysis

Until a few years ago, the default workflow in science and industry was to decide on a particular (research) question, select which of the established methods to use to address this question, and then select a specific collection and sampling strategy to acquire data that will fit the needs established by the methods and research questions. Sometimes this involved inventing new methods and observing how they would perform when applied to the data, or, the other way around, realize that the collected data could be better explored using a new method. Not all data was collected from scratch, researchers always shared data or used some common base repositories.

Today, however, we often see this typical workflow reversed. Nowadays, there is an abundance of data at an ever increasing spatial, temporal, and thematic resolution. Together with a wide set of established methods and computational cyber-infrastructures, it becomes possible to start with the data first and ask research questions after mining the data for patterns and uncovering correlations. To give a concrete example, the majority of research on online social networks starts with data analysis and exploration, i.e., the research questions to be asked are motivated by the existence of particular datasets. Even for established workflows, it is increasingly common to search for suitable data across repositories and disciplines and combine them. Finally, an increasing amount of data is contributed by citizens (Elwood, Goodchild, and Sui 2013).

While the difference between collecting own data and using data collected by others seems small, it is, in fact, a major change that impacts how science works. With the higher resolution of the data, nuanced differences become more important and with the wide variety of sources, the quality and provenance of these data is more difficult to control. In the data age, synthesis becomes the new analysis and the ability to find, reuse, and integrate data from multiple, highly heterogeneous sources on-the-fly becomes a major aspect



Federal datasets are subject to the U.S. Federal Government Data Policy. Non-federal participants (e.g., universities, organizations, and tribal, state, and local governments) maintain their own data policies. Data policies influence the usefulness of the data.

Map data CC-BY-SA by OpenStreetMap
Tiles by MapQuest

Natural Disaster

93 datasets found for "Natural Disaster"

Datasets ordered by Relevance

You are searching in the list of datasets. Show results in entire Data.gov site.

geospatial (62)
non-geospatial (31)
earth science (49)
disasters (48)

FEMA Disaster Declarations Summary

Federal Emergency Management Agency, Department of Homeland Security — FEMA Disaster Declarations Summary is a summarized dataset describing all federally declared disasters. This information begins with the first disaster declaration...

XLS json, xml, bson

Noninsured Crop Disaster Assistance Program

Department of Agriculture — USDA's Farm Service Agency's (FSA) Noninsured Crop Disaster Assistance Program (NAP) provides financial assistance to producers of noninsurable crops when low...

HTML

Child Nutrition Programs Disaster Response Memo

Department of Agriculture — This memorandum provides an overview of ways State agencies, School Food Authorities (SFA) participating in the National School Lunch and School Breakfast...

Federal datasets are subject to the U.S. Federal Government Data Policy. Non-federal participants (e.g., universities, organizations, and tribal, state, and local governments) maintain their own data policies. Data policies influence the usefulness of the data.

Map data CC-BY-SA by OpenStreetMap
Tiles by MapQuest

Earthquakes

243 datasets found for "Earthquakes"

Datasets ordered by Relevance

You are searching in the list of datasets. Show results in entire Data.gov site.

geospatial (224)
non-geospatial (19)
earth science (148)
noaa (134)

Earthquake Feeds

U.S. Geological Survey, Department of the Interior — Near real-time earthquake information for a variety of time windows in a variety of formats.

csv, kml

Earthquake Locations

State of North Dakota — This layer has been compiled from various sources depicting the locations of earthquakes that have been confirmed to have occurred within the state of North Dakota.

HTML HTML WMS Esri REST Esri REST

Earthquake Damage - General

National Oceanic and Atmospheric Administration, Department of Commerce — An earthquake is the motion or trembling of the ground produced by sudden displacement of rock in the Earth's crust. Earthquakes result from crustal strain,...

HTML HTML

Figure 2: Searching Data.gov for natural disaster datasets.

of modern science and business. For instance, according to the NASA, scientists and engineers spend more than 60% of their time just preparing data for model input or data-model intercomparison (NASA 2012). Complex scientific and societal questions cannot be answered from within one domain alone, but involve multiple disciplines. For instance, studying the relation between natural disasters and migration to improve policy making requires data and models from a wide range of domains including economics, political sciences, medical sciences, epidemiology, geography, geology, climatology, demographics, and so forth.

A simple example can help to illustrate the difficulty in retrieving relevant information. Data.gov is the main aggregator through which the US government makes its open data publicly available . It is arguably the biggest data repository on the Web. Searching this portal for data about natural disasters will return 93 datasets, while searching for a specific kind of disaster, namely earthquakes, will return 243 results; see figure 2. Making data discoverable and integrable remains a major challenge.

Research areas such as semantic interoperability, symbol grounding, cyber-infrastructures, eScience, Web science, and many others, try to address these problems from different viewpoints. Semantic interoperability, for instance, asks the question whether a sequence of processing steps which accept each others outputs as inputs will produce meaningful results. In contrast, Web science and information observatories aim at monitoring and understanding the Web and its social machines (Hendler et al. 2008). In the following we discuss how the Semantic Web contributes to addressing these challenges.

Smart Data versus Smart Applications

A major paradigm shifts introduced by the Semantic Web is to focus on the creation of smart data instead of smart applications. The rationale behind this shift is the insight that smart data will make future applications more (re)usable, flexible, and robust, while smarter applications fail to improve data along the same dimensions (Janowicz and Hitzler 2012). As a result, Semantic Web applications and infrastructures can be put together largely by using of-the-shelf software such as faceted browsing frameworks and user interfaces, transformation, alignment, and link discovery tools, reasoners, editors, and a wide variety of programming frameworks. Even more, in most cases using and combining these software only requires minimum effort. For instance, instead of having to develop graphical user interfaces to explore a new dataset, faceted browsing frameworks load the used ontologies to extract and populate the facets. Similarly, in terms of maintenance, changes in the underlying ontologies and the used data can be automatically reflected by the user interface. Due to the early, open, and rigid W3C standardization process, combining Semantic Web software and moving data between them is largely painless. Finally, a majority of the tools and frameworks are available as free and open source software. This is especially important for researchers in developing countries and also ensures the longevity of the Semantic Web as infrastructure.

Interestingly, this is in clear contrast to the evolving big data landscape. As the basic assumption behind big data is that the sheer availability of data at an ever increasing spatial, temporal, and thematic resolutions will be a game changer, the majority of work focuses on complex and largely commercial systems for predictive analytics, i.e., transforming data into insights. The involved (industry) players compete for customers and common standards have not yet been established. This is also true for many of the cyber-infrastructures developed by the researcher community. While most of them share the same vision of online workbenches in which data can be loaded, manipulated, and analyzed in the Cloud, most systems do not address data interchange and interoperability.

The difficulty with this strategy and investing heavily in searching the needle in the haystack instead of cleaning up the haystack first, is that the data remains unaltered and has to be cleaned, classified, analyzed, and annotated, over and over again. While this makes sense for large amounts and high velocities of unstructured data such as tweets, there is much to gain by semantically annotating and sharing data according to some well established vocabularies, e.g., Schema.org. Once the data is semantically lifted, it can be more easily discovered, reused, and integrated (Hendler 2013).

Vocabulary Diverse Data

One may note that transforming data into Linked Data or semantically annotating them introduces a particular viewpoint or context and that this hinders the free re-usability and analysis of data. For instance, categorizing a relationship between two individuals as *marriage* enforces a particular socio-cultural viewpoint that may not be accepted in other societies. This is certainly a valid concern but it ignores three key aspects.

First, there is no such thing as *raw* data. Data is always created for a certain purpose, following workflows and observation procedures, depends on the used sensors and technologies, comes with an intrinsic uncertainty, reflects the theories and viewpoints of the people that recorded the data, and so forth. To give a concrete example, the body position at which a blood pressure measure was taken matters for the interpretation of the results. Therefore, it is important to agree on a terminology for body positions. SNOMED CT, for instance, provides a *Body position for blood pressure measurement* term, while the National Cancer Institute Thesaurus provides definitions for the positions *sitting*, *standing*, and *supline*. Blood pressure observations can more easily be interpreted and combined meaningfully if these (meta)data are present. Ontologies and Semantic Web Technologies support the (semi-automatic) annotation of data and are thus heavily applied in domains such as the life sciences (Goldfain et al. 2013). Consequently, semantic annotations do not restrict the use of the data but make existing usage restrictions explicit.

Second, restricting the interpretation of data towards their intended meaning is one of the goals of the Semantic Web. With respect to the marriage example above, one cannot simply collect and compare the percentage of married couples per state in the US to arrive at any meaningful conclusion without taking into account that states such as California allow for same-sex marriage while Florida does not. To make things even more complicated, states that do allow for same-sex marriage do not necessarily recognize the marriages performed in other jurisdictions. Finally, states such as Oregon recognize same-sex marriages from other jurisdictions; however, they do not allow them to be performed within their own territory. Ontologies make these differences *explicit* and allow Semantic Technologies to exploit them, e.g., to determine whether data can be integrated or not.

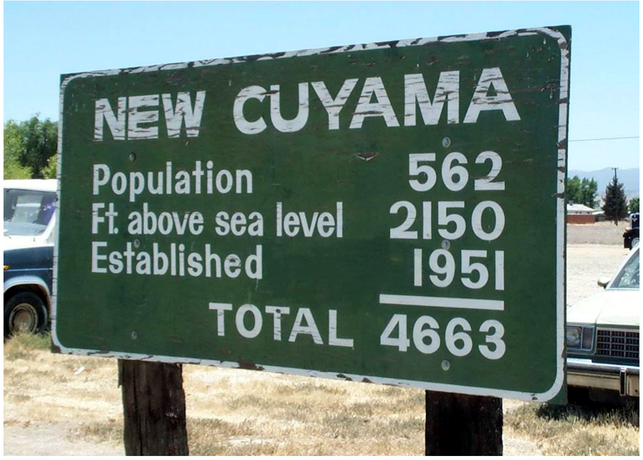


Figure 3: A sign in New Cayama, Santa Barbara County, California; original picture by Mike Gogulski (CC BY 2.5).

It is worth mentioning that one of the factors for the success of machine learning in the big data field is the crowdsourcing of data labeling, e.g., via Amazon's Mechanical Turk. The downside of this approach, however, is that only the default meaning is captured and important domain nuances are lost. For instance, studying and addressing the problem of deforestation cannot be done without taking into account the hundreds of different and often legally binding definitions of *forest* (Lund 2014 rev). As illustrated by Sasaki and Putz, subtle changes to how concepts such as forest are defined may have major social and physical impacts (Sasaki and Putz 2009). Ontologies make such differences explicit.

With respect to data science and predictive analytics, ontologies and Semantic Technologies can also restrict the operations that should be performed on a given dataset. A common example used to make this point is a sign in New Cayama, Santa Barbara County, California; see figure 3. While Stevens' scales of measure clearly permit addition, the operation is not meaningful given the semantics of the observable properties population, elevation, and the year New Cayama was established (Compton et al. 2012).

Third, the same data can be described via several vocabularies and those vocabularies can be interrelated. The fact that a dataset about important historic figures born in London, UK was described using the BBC core ontology does not mean that the very same dataset cannot also be described using Schema.org. Additionally, those vocabularies can be put in relation. The latter explicitly supports undead persons while the BBC core ontology does not make any statements about fictional persons. Thus, one could axiomatically state that the Schema.org perspective is more inclusive than the BBC core vocabulary.

To give a more sophisticated example, the BBC and Schema.org vocabularies both allow to specify the occupation (job title) of a person. However, they do not allow to specify a duration. A more complex ontology may use concept reification to introduce a temporal scope for occupations. These different models can co-exist, be used to describe the same data, and even be aligned. Thus, the degree of formalization and *ontological commitments* can be tailored to particular application or analysis needs.

Big Data, Small Theories

In fact, such alignments are a key feature of the Semantic Web. In contrast to early work on information ontologies, the Semantic Web does not require to agree on a specific and limited set of well-defined and coherent ontologies. Consequently, data publishers can define their own local ontologies, reuse existing ontologies, or combine multiple ontologies. Strictly speaking, the Semantic Web stack does not impose *any* restrictions on the use and quality of ontologies. For instance, one can describe data using only selected predicates from different ontologies and add additional, even inconsistent, axioms. While this would not be considered good practice and will prevent the usage of certain (reasoning) capabilities of Semantic Web Technologies, basic queries will still be possible.

This is not a design flaw. To the contrary, it is a necessary feature of Web-scale, heterogeneous knowledge infrastructures that embraces the AAA slogan (Allemang and Hendler 2011) that **A**nyone can say **A**nything about **A**ny topic (at **A**ny place at **A**ny time, to add two more As). Together with the diversity aspects discussed above, this moves the Semantic Web and graph-databases closer to the NoSQL realm than to the traditional relational database landscape. More importantly, however, is allows for a purpose-driven and staged approach in which the level of semantics and the choice of ontology depends on the application needs. By implementing the Open World Assumption, the formal semantics of knowledge representation languages such as the Web Ontology Language is tailored to the decentralised structure of Web-scale infrastructures in which statements can disagree and recourses can be temporarily unavailable. The truth of a statement, e.g., whether Crimea belongs to the Ukraine, is independent of whether it is known to the system or not. Therefore, and in contrast to most databases, a lack of such statement does not imply that it is false. Statements on the Semantic Web can also be contradictory without causing the whole system to fail. This follows a well established position in artificial intelligence research, namely to be locally consistent while allowing for inconsistencies on the global knowledge base level (Wachsmuth 2000). For instance, a Linked Data hub such as DBpedia will store one day of birth for a given person while there may be multiple alternative birthdays available on the overall Linked Data cloud.

As the Semantic Web provides alignment methods that allow to relate multiple ontologies, it supports federated queries and synthesis across a variety of highly heterogeneous data sources without enforcing global agreement nor losing control over the own data and schemata (Noy and Musen 2000; Cruz, Antonelli, and Stroe 2009; Jain et al. 2010; David et al. 2011). This allows data providers and users to individually decide on their ontology of choice, the level of required and meaningful axiomatization, and relation to other ontologies. As a result, we are seeing an increasing amount of interrelated but highly specialized small ontologies being developed that are used in practice instead of having to wait for globally agreed schemata to publish, retrieve, reuse, and integrate data. At the same time, the needs for a lightweight semantic annotation is served by Schema.org, the DBpedia ontology, and so forth.

Linking and Exploring Data

One of the key characteristics introduced by Linked Data is the idea of assigning URIs as globally unique identifiers for information resources, e.g., Web pages, as well as for non-information resources, e.g., sensors and people (Berners-Lee 2006). Dereferencing URIs for non-information resources should lead to information about these resources, e.g., RDF triples about a person. In addition, Linked Data proposes to interlink those URIs to foster discovery, e.g., by learning about the birth place of the person and then exploring which other people were born there, learning about events they were involved, and where those events took place. Focusing on the *relations* between objects, actors, events, and places, that jointly form a global graph seems in many ways more promising than a Big Data cloud that is merely a collection of isolated facts. This follow-your-nose approach is central to the exploration of unfamiliar sources for which it is difficult to pose an exact search query. Exploratory search (Marchionini 2006) in general is considered to be more appropriate for navigating large infrastructures of highly heterogeneous data than classical information retrieval and querying.

Now establishing such links would be a very labor intensive task and infeasible to perform manually for high volume and high velocity data. To address this issue, the Semantic Web offers an increasing number of frameworks that can automatize the discovery and establishment of links. The SILK link discovery engine is one such system (Volz et al. 2009) and also offers an intuitive Web-based online workbench to specify and execute linkage rules.

While many different types of links can be established between resources, e.g., linking a patient's record to a particular disease, identity links are considered to be especially important as they play a key role in co-reference resolution, conflation, and de-duplication. These *sameAs* links declare two resources identified by different URIs to be the same resource. Note that this is a stronger statement than declaring resources to be equal or similar. For instance, two replicas/copies of a painting can be indistinguishable but are not the same. Thus, if two resources are linked via a sameAs relation, every statement about one of them also applies to the other resource.

It is interesting to note how inductive and deductive methods play together to foster data retrieval, reuse, and integration. Top-down engineered ontologies and logical inferencing play a key role in providing the vocabularies for querying data, while machine learning techniques enable the linkage of these data. Increasingly, however, inductive and deductive methods are also combined to mine knowledge patterns and ontological primitives to scale up the development of ontologies (Gangemi and Presutti 2010).

URIs as global primary keys and identity links jointly form a powerful framework for the de-silofication of local databases and therefore are a major building blocks for next-generation knowledge infrastructures.

Compressing and Maintaining Data

Unsurprisingly, more data does not automatically translate into more insight. Nonetheless, data with a higher spatial, temporal, and thematic resolution can be used to gain new insights and make better predictions. While answers to frequent queries may be materialized, most data are rarely requested and can be compressed to reduce volume. Interestingly, Semantic Technologies and ontologies are very suitable for data compression. In case of Linked Data, for instance, instead of storing each triple, one only has to store such triples that cannot be inferred from other triples via an ontology. To give a simple example, by declaring *partOf* to be transitive, one does not have to explicitly store the fact that any given city is part of a certain county, state, and continent. Storing the facts that a city is part of a county, the county is part of a state, and so forth, is sufficient. Similarity, by using *role chains* a family ontology can define that the parents of a person's parent are this person's grandparents. Thus, this fact does not need to be stored in the data. A few of these axioms can drastically reduce the amount of data that has to be stored and downloaded (Joshi, Hitzler, and Dong 2013).

Even more importantly, ontologies also help to maintain datasets. Instead of having to update a multitude of records to adjust to new or changing information, only those statements have to be updated that cannot be inferred via a background ontology. To use the marriage example introduced before, instead of checking tens of thousands of records every time the law changes in one of the US states, a few updates to the used ontology are sufficient to reflect the changes and to infer which of the stored marriage relations have to be updated.

Combining Inductive and Deductive Methods

In principle, the dedicated pursuit of combining inductive and deductive techniques for dealing with data may indeed have significant potential which remains to be unlocked, and the key would be in a best-of-both-worlds combination. Indeed, deductive methods are extremely powerful if used in special-purpose applications such as expert systems, with a well-defined use case, limited scenario, and a domain which is understood well enough so that expert knowledge can be captured in the form of crisp logical axioms. Inductive methods, on the other hand, excel if data is noisy, expert knowledge is not readily mapped, and the input-output relationship matches the search space, i.e., can be captured by learning with the chosen method(s). Yet, deductive methods usually break down under noisy data, while inductive methods may solve a problem but may not help to understand the solution or to verify it.

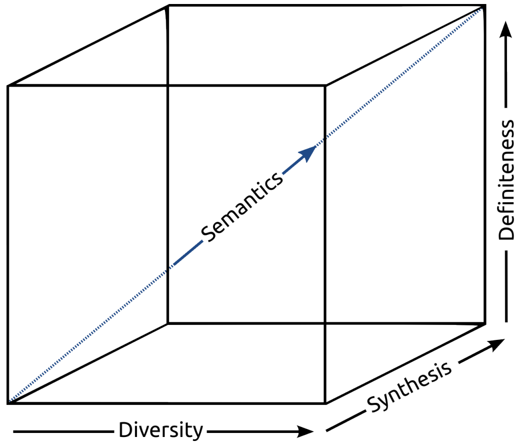


Figure 4: The semantic cube depicts how the need for Semantic Technologies and ontologies raises with an increasing role of diversity, synthesis, and definiteness.

It has been argued that much benefit may lie in a more systematic study of combinations of deductive and inductive methods. For example, systems which first learn higher-level features, expressed as logical axioms, from data, and then use these higher level features for non-trivial deductive inferences, are few. Another perspective would be that of viewing a deduction problem as an information retrieval or binary classification problem, and then applying non-deductive techniques to solve this task (Hitzler and van Harmelen 2010). Little research has so far been undertaken along this line.

Conclusions

While data-intensive science and predictive analytics take the existence of data at an ever increasing spatial, temporal, and thematic resolution as a given. Their success ultimately depends on whether these data can be retrieved, accessed, reused, and integrated in a meaningful way. Given the volume, velocity, and especially variety of today's data universe it is increasingly difficult to understand whether a certain dataset is fit for a given task/model and whether different datasets can be combined. The seemingly same terminology often has a different meaning in another domain. Even within a single domain the improved resolution comes at the costs of increasingly nuanced differences in the used methods, models, and terminologies. Ignoring those differences affects the quality and reproducibility of research results or market analysis in the industry. In this paper we discussed how the Semantic Web can address some of these challenges and why it is a suitable, robust, affordable, open, and persistent infrastructure for data-intensive science. We illustrated our arguments with various examples from different scientific domains.

The discussed value proposition of the Semantic Web for data-intensive science and industry applications can be best summarized by depicting the increasing need for semantics along a number of dimensions. Figure 4 illustrates this by introducing a *semantic cube*.

- The **diversity dimensions** includes aspects such as the degree of interdisciplinarity, the semantic heterogeneity of the data, the variety of involved media formats, perspectives, and themes, as well as their varying resolution, uncertainty, lineage, and credibility. Consequently, this dimension is focused on *data*. An increasing diversity leads to an increasing need for formal semantics, e.g., to mediate between different terminologies or in case of different viewpoints.
- The **synthesis dimension** represents the difference in the performed *tasks*. An increasing need for data reuse, integration, conflation, and synthesis, leads to an increasing need for semantics. In contrast, tasks such as computing descriptive statistics for a single numerical dataset benefit less from Semantic Technologies and ontologies. Semantic Web Technologies play a role in preparing data for analysis, e.g., by ensuring that the retrieved data is suitable as input for a certain model, as well as in sharing the analysis results, i.e., as communication layer that makes the meaning of the resulting data explicit.
- The **definiteness dimension** represents aspects of *purpose* and intended use. For instance, while machine learning is concerned with a degree of certainty, Semantic Technologies aim at logical consequences. To give a concrete example, reasoning services such as instance checking or concept subsumption produce guaranteed, i.e., rigid, results. If, by definition, all rivers flow into other waterbodies (such as oceans, lakes, or other rivers) and the Rhein flows into the North Sea, then the North Sea must be a waterbody. One could now argue that some rivers flow into the ground or dry up entirely without reaching another body of water. This, however, should not be misunderstood as invalidating the inference made above. Ontologies define a data provider's *view* on these data, they do not define some sort of canonical *reality*. By making the provider's perspective explicit, ontologies reduce the risk of misunderstanding and misusing data.

Consequently, the need for and benefit from using Semantic Technologies increases as those three dimensions increase, i.e., with an increasing data and domain diversity, with an increasing role of data synthesis, and with an increasing importance of definiteness.

Acknowledgement. Pascal Hitzler acknowledges support by the National Science Foundation under award 1017225 "III: Small: TROn – Tractable Reasoning with Ontologies." Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Krzysztof Janowicz is an assistant professor for geoinformatics at the University of California, Santa Barbara, USA.
Frank van Harmelen is a professor in knowledge representation and reasoning at the Vrije Universiteit Amsterdam, Netherlands.
James A. Hendler is a professor for artificial intelligence at the Rensselaer Polytechnic Institute, USA.
Pascal Hitzler is an associate professor for Semantic Web technologies at the Wright State University, USA.

References

Allemang, D., and Hendler, J. 2011. *Semantic web for the working ontologist: effective modeling in RDFS and OWL*. Elsevier.

Berners-Lee, T. 2006. Linked data: Design issues. http://www.w3.org/DesignIssues/LinkedData.html.

Compton, M.; Barnaghi, P.; Bermudez, L.; Garca-Castro, R.; Corcho, O.; Cox, S.; Graybeal, J.; Hauswirth, M.; Henson, C.; Herzog, A.; Huang, V.; Janowicz, K.; Kelsey, W. D.; Phuoc, D. L.; Lefort, L.; Leggieri, M.; Neuhaus, H.; Nikolov, A.; Page, K.; Passant, A.; Sheth, A.; and Taylor, K. 2012. The SSN ontology of the W3C semantic sensor network incubator group. *Web Semantics: Science, Services and Agents on the World Wide Web* 17(0):25 – 32.

Cruz, I. F.; Antonelli, F. P.; and Stroe, C. 2009. Agreementmaker: efficient matching for large real-world schemas and ontologies. *Proceedings of the VLDB Endowment* 2(2):1586–1589.

David, J.; Euzenat, J.; Scharffe, F.; and Trojahn dos Santos, C. 2011. The Alignment API 4.0. *Semantic Web* 2(1):3–10.

Dropp, K.; Kertzer, J. D.; and Zeitzoff, T. 2014. The less americans know about Ukraine's location, the more they want U.S. to intervene. http://www.washingtonpost.com/blogs/monkey-cage/wp/2014/04/07/the-less-americans-know-about-ukraines-location-the-more-they-want-u-s-to-intervene/.

Elwood, S.; Goodchild, M. F.; and Sui, D. 2013. Prospects for VGI research and the emerging fourth paradigm. In *Crowdsourcing Geographic Knowledge*. Springer. 361–375.

Gangemi, A., and Presutti, V. 2010. Towards a pattern science for the semantic web. *Semantic Web* 1(1):61–68.

Goldfain, A.; Xu, M.; Bona, J.; and Smith, B. 2013. Ontology based annotation of contextualized vital signs. In Dumontier, M.; Hoehndorf, R.; and Baker, C. J. O., eds., *Proceedings of the 4th International Conference on Biomedical Ontology 2013*, volume 1060 of *CEUR Workshop Proceedings*, 28–33. CEUR-WS.org.

Hendler, J.; Shadbolt, N.; Hall, W.; Berners-Lee, T.; and Weitzner, D. 2008. Web science: an interdisciplinary approach to understanding the web. *Communications of the ACM* 51(7):60–69.

Hendler, J. 2013. Peta vs. meta. *Big Data* 1(2):82–84.

Hey, A. J.; Tansley, S.; Tolle, K. M.; et al. 2009. The fourth paradigm: data-intensive scientific discovery. Microsoft Research Redmond, WA.

Hitzler, P., and van Harmelen, F. 2010. A reasonable semantic web. *Semantic Web* 1(1):39–44.

Jain, P.; Hitzler, P.; Sheth, A. P.; Verma, K.; and Yeh, P. Z. 2010. Ontology alignment for linked open data. In *The Semantic Web–ISWC 2010*. Springer. 402–417.

Janowicz, K., and Hitzler, P. 2012. Key ingredients for your next semantics elevator talk. In Castano, S.; Vassiliadis, P.; Lakshmanan, L. V.; and Lee, M.-L., eds., *Advances in Conceptual Modeling – ER 2012 Workshops CMS, ECDM-NoCoDA, MoDIC, MORE-BI, RIGiM, SeCoGIS, WISM, Florence, Italy, October 15-18, 2012. Proceedings*, volume 7518 of *Lecture Notes in Computer Science*, 213–220. Springer.

Joshi, A. K.; Hitzler, P.; and Dong, G. 2013. Logical linked data compression. In Cimiano, P.; Corcho, Ó.; Presutti, V.; Hollink, L.; and Rudolph, S., eds., *The Semantic Web: Semantics and Big Data, 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013. Proceedings*, volume 7882 of *Lecture Notes in Computer Science*, 170–184. Springer.

Lund, G. 2014 rev*. Definitions of forest, deforestation, afforestation, and reforestation. Technical report, Forest Information Services.

Marchionini, G. 2006. Exploratory search: from finding to understanding. *Communications of the ACM* 49(4):41–46.

NASA. 2012. A.40 computational modeling algorithms and cyberinfrastructure (December 19, 2011). Technical report, National Aeronautics and Space Administration (NASA).

Noy, N. F., and Musen, M. A. 2000. Algorithm and tool for automated ontology merging and alignment. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-00). Available as SMI technical report SMI-2000-0831*.

Sasaki, N., and Putz, F. E. 2009. Critical need for new definitions of forest and forest degradation in global climate change agreements. *Conservation Letters* 2(5):226–232.

The Economist. 2012. Indian, Pakistani and Chinese border disputes: Fantasy frontiers. http://www.economist.com/blogs/dailychart/2011/05/indian_pakistani_and_chinese_border_disputes.

Volz, J.; Bizer, C.; Gaedke, M.; and Kobilarov, G. 2009. SILK – a link discovery framework for the web of data. In Bizer, C.; Heath, T.; Berners-Lee, T.; and Idehen, K., eds., *Proceedings of the WWW2009 Workshop on Linked Data on the Web, LDOW 2009, Madrid, Spain, April 20, 2009*, volume 538 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Wachsmuth, I. 2000. The concept of intelligence in AI. In *Prerational Intelligence: Adaptive Behavior and Intelligent Systems Without Symbols and Logic, Volume 1, Volume 2 Prerational Intelligence: Interdisciplinary Perspectives on the Behavior of Natural and Artificial Systems, Volume 3*. Springer. 43–55.