Semantic Web 3 (2012) 213–221
DOI 10.3233/SW-2012-0070
IOS Press

Editorial

The Digital Earth as knowledge engine

Krzysztof Janowicz [a] and Pascal Hitzler [b]
[a] *University of California, Santa Barbara, USA*
E-mail: jano@geog.ucsb.edu
[b] *Kno.e.sis Center, Wright State University, Dayton, OH, USA*
E-mail: pascal.hitzler@wright.edu

Abstract. The *Digital Earth* [13] aims at developing a digital representation of the planet. It is motivated by the need for integrating and interlinking vast geo-referenced, multi-thematic, and multi-perspective knowledge archives that cut through domain boundaries. Complex scientific questions cannot be answered from within one domain alone but span over multiple scientific disciplines. For instance, studying disease dynamics for prediction and policy making requires data and models from a diverse body of science ranging from medical science and epidemiology over geography and economics to mining the social Web. The naïve assumption that such problems can simply be addressed by more data with a higher spatial, temporal, and thematic resolution fails as long as this *more* on data is not supported by more knowledge on how to combine and interpret the data. This makes semantic interoperability a core research topic of data-intensive science. While the Digital Earth vision includes processing services, it is, at its very core, a data archive and *infrastructure*. We propose to redefine the Digital Earth as a *knowledge engine* and discuss what the Semantic Web has to offer in this context and to Big Data in general.

> *'Considerable data regarding the environment are available through the myriad of remote-sensing programs, however, this data is not directly compatible with the models. It has been observed that scientists and engineers spend more than 60% of their time just preparing the data for model input or data-model intercomparison. This is an inefficient use of the precious time of NASA scientists and engineers.'* [28]

Beyond the general-purpose Web

Initially, the Semantic Web [3,17] was proposed as a successor of the document Web that makes the stored content understandable to software agents and enables them to extract, process, and combine this information. At this time, the Web was still dominated by authoritative sources and different from the social read-write Web that we know today. During these early days, data on the Web was assumed to be relatively stable, authoritative, and fit for a given, predefined purpose. Thus, in analogy to catalogs, it was assumed that data providers would invest in creating intelligent metadata to improve retrieval and reuse. This made semantic technologies capable of handling sophisticated ontologies a promising research vision.

These days, however, the Web is based on fundamentally different principles. The *volume* of data is growing at a higher rate than our capacities for long-term archiving. New data is added at a *velocity*, surpassing our ability to consume it. Instead of a limited number of data providers and formats, data is contributed by a myriad of human users, software agents, and technical sensors in a *variety* of different multimedia formats. While these three *V*'s are characteristic for the omnipresent Big Data, we argue that a fourth *V* addressing the *value* of the created data is relevant as well. Finally, the general-purpose Web in itself is losing ground with traffic constantly declining since more than 10 years and the increasing success of single-purpose *apps* [1].

These new realities call for a new perspective on the vision of a semantic Web. Central assumptions such as that data providers would invest into creating sophisticated and stable ontologies do not hold any longer. To the contrary, the Web calls for pattern-like, application-driven schema knowledge that can be easily adopted and makes data reusable for use cases not envisioned by the data provider beforehand. In fact, Linked Data [4] can be understood as such a novel view on Web semantics. It takes up the successful idea of Web links and enriches them with names and types, proposes a decentralized and open network of interlinked data providers, makes data uniquely referenceable using URIs, and argues for lightweight ontologies. Schema.org, a schema collection for embedded semantic markup for Web pages, launched by Google, Yahoo, and Microsoft in 2011 points into a similar direction. Even more, Google's Knowledge Graph implements such a lightweight version of a semantics-based search on Web scale.

Is there no need for more powerful knowledge representation and reasoning? There is, but we have to realize that different settings require different methods. We have to look beyond the general-purpose Web to communities that have an intrinsic need for more intelligent metadata and conceptual modeling. In terms of the Big Data *V*'s, volume and velocity in themselves do not motivate the need for more elaborate semantic technologies, but may, indeed, be best served with lightweight approaches. It is the variety and value dimension of data that motivates the investment into technologies for modeling and reasoning over complex knowledge. In the following we argue that interdisciplinary science offers just this setting, in which heterogeneous data is painstakingly created, collected, maintained, and integrated to answer complex scientific and social questions, and to support policy making.

The Digital Earth

Introduced by Al Gore in 1998 [13] and refined over the years, the Digital Earth envisions a highly interdisciplinary knowledge archive and service infrastructure for geo-referenced, interconnected, multi-dimensional, multi-thematic, and multi-perspective data [8]. The 2011 symposium of the International Society for Digital Earth, for example, was themed *The Knowledge Generation* and investigated the role of Digital Earth technologies for economic and social sustainable development, disaster mitigation, environmental protection, conservation of natural resources, as well as the improvement of living standards.[1] The Digital Earth is motivated by the insight that complex scientific questions cannot be answered from within one domain alone but span over multiple scientific disciplines ranging from the natural and earth sciences to the social sciences, information science, and engineering. Essentially, the Digital Earth is about the exchange, integration, and reuse of heterogeneous data. This makes semantic interoperability a major research topic. Over the years some of the initial Digital Earth goals have been realized by virtual globes such as NASA World Wind or Google Earth. However, these solutions are mostly focused towards visualization and simple retrieval tasks.

Instead of establishing interoperability by sacrificing semantic heterogeneity, the Digital Earth calls for methods to reason in the presence of heterogeneous and contradicting conceptual models, while maintaining the variety brought in by different scientific domains. A key problem in exchanging scientific data and using such data for policy making is that the meaning of categories used to share knowledge is not made explicit. Hence, the same terms have radically different meanings. This is especially troublesome for terms that seem to be part of common everyday language. Consequently, they are often not defined when publishing scientific data. Typical examples include *city*, *forest*, or *boundary* [23]. Moreover, the theories and methods used to produce these categories are most often not shared together with the data which limits our ability to reproduce scientific results [7].

In the past, description logics-based knowledge representation languages have often been mistaken as a replacement to numerical and statistical modeling. Instead, ontologies are best understood as a thin communication and exchange layer. Thin, in this context, should not be confused with unimportant; in fact, ontologies are the decisive glue between models, data, and users. Ontologies should assist in answering questions such as whether a specific model can be meaningfully applied to a particular dataset and whether this dataset is compatible with the user's conceptualization of the domain at hand. For instance, ontologies and reasoning systems could assist users in selecting study areas and datasets to test their scientific hypotheses.

[1] http://www.isde7.net/

We argue that in the light of IBM's DeepQA architecture [10] and recent progress on Semantic Web, Linked Data [4], and geospatial semantics [25], the Digital Earth should be envisioned as a distributed *knowledge engine* and question answering system that supports scientists beyond mere data retrieval. While the need for ontologies and semantic technologies is widely acknowledged, crucial components to realize such a vision are missing: how to assist scholars in defining micro-ontologies that support the conceptualization of their local models, how to arrive at the primitives, i.e., base symbols, for such ontologies, how to ground primitives in real observations and align them to knowledge patterns, how to track categorical data back to measurements using provenance, how to make ontologies first class citizens of statistic methods, and, finally, how to reason over heterogeneous, incomplete, and contradicting micro-ontologies to foster interoperability and for checking integrity constrains before reusing data [23]? In other terms, how do we enable domain scientists to become knowledge engineers and at the same time keep the underlying Semantic Web machinery transparent?

Sources of variety

The variety of Big Data in general and the Digital Earth in specific stems from different sources.

First, and most obviously, different scientific disciplines use the same terms while the underlying meaning often differs to a degree where they become incompatible. A good example is the use of the term *scale* in geography versus most other sciences. Intuitively, a *large scale* study in, say, ecology or economy, covers a large extent in surveyed space; where the notion of a space is not restricted to its spatial meaning but includes attribute spaces as well. Modeling the global economic impact of the 2009 flu pandemic based on public heath data, for instance, would be called a large scale study.

In contrast, geography is following the traditional cartographic definition of scale. Here, scale is the representative fraction between the distance on a map and the corresponding distance on the ground. For example, while each unit on a 1/24 000 map represents 24 000 units, e.g., centimeter, on the ground, each unit on a 1/500 000 map corresponds to 500 000 units on the ground. As the first fraction is greater than the second, a 1/24 000 map is a large scale map covering a small extent of the Earth's surface in detail. To the contrary, a small scale covers large areas, such as continents, at the cost of representing less details.

While the example just given is multi-thematic, streets are often used to illustrate multiple perspectives on geographic space and categorization. A street is a *connection* from A to B from the view point of transportation science, while it is a disruptive *separation* that cuts a habitat into segments from the view point of ecology and conservation.

Differences in the meaning of the used terms are even more troublesome when they happen within scientific communities as a common agreement is often wrongly assumed. Meaning is not static but dynamically reconstructed during language use. While humans can perform this reconstruction by situated simulation [2,15], terms used in metadata records are static and de-contextualized. Consequently, the challenge is to understand what was meant with a keyword used to annotate data many years ago [30]. As Scheider puts it *the problem is not that machines are unable to communicate, but that humans misunderstand each other if communicating via machines* [29]. Finally, meaning does not only vary across and within scientific communities, but also as a function of language, space, culture, age, social structure, and many other factors.

A second and related source for variety lies in the very nature of knowledge itself. While some scientists favor a Platonic realism and argue for an independent ontological existence of universals, this position is difficult to defend in case of highly multidisciplinary research that cuts through the boundaries of social and natural sciences. Acknowledging that the classes we define in our ontologies are constructed and that knowledge is an evolving process of adaptation to our experiential reality [31] implies that there is (and will be) more than just one way to construct. Consequently, Big Data should not be approached with equally big theories that try to arrive at a universal and static agreement, but by a network of theories that foster interoperability without giving up on semantic heterogeneity (and, thus, the long-tail of science).

A third reason for variety in Big Data is what could be called an observation versus definition mismatch. For example, a transportation infrastructure ontology may model watercourses in terms of water depth, currents, and hazard to navigation, while another source may model watercourses by their Strahler number, i.e., by their branching complexity. While both models are useful and may be applied to describe the same entities on the surface of the Earth, the observation data of the first one cannot be transformed to match the second definition and vice versa. While this is related to the multi-purpose argument made before, it adds the additional problem that the definition of terms cannot be reconstructed out of the available observation data.

Finally, another source of variety stems from the way how data and conceptual models are produced. This ranges from different scientific workflows and measurement procedures to different cultures and file formats. To give a concrete example, authoritative providers such as the U.S. Geological Survey (USGS) aim at a high degree of standardization, a stable schema level, maintainable data products, and well defined measures for data quality. In contrast, so-called Volunteered Geographic Information [11] such as known from OpenStreetMap, Ushahidi, and georeferenced Flickr images and tweets, are created and maintained by a highly-heterogeneous user community with different backgrounds and application areas in mind. This kind of citizen science, which is also very popular in other scientific domains, relaxes the typical rules under which data is collected for the benefit of providing the most up-to-date data. From the viewpoint of Digital Earth research this source of variety opens up new possibilities for science and especially for the evaluation of data. For instance, Flickr images can be used to validate tweets about the Arab Spring, and volunteered crisis mapping can show a complementary picture of the 2011 Earthquake in Japan.

Research challenges

In the previous sections, we argued that a distributed knowledge engine that cuts through scientific domains may be a promising vision for the next decade of semantics research. We explained why the variety and value dimensions of Big Data will benefit most from semantics research which enables a more efficient publishing, retrieval, reuse, and integration of scientific data, models, and methods. In the following, we highlight selected research challenges that would have to be addressed to realize the vision of the Digital Earth as a knowledge engine.

Fields and objects

Data can typically be represented as fields or as objects [12]. For example, terrain can be modeled as a continuous surface of elevation values or by discrete objects, e.g., hills and mountains. In scientific workflows, sensor data is often collected as fields and transformed into objects later (if at all) during analysis or information visualization. A typical example is the classification of continuous absorption and reflection patterns of electromagnetic radiation collected by remote sensing instruments into discrete land cover classes. The classed data is often shared as objects, e.g., polygons representing *forests*. Semantic Web technologies and the methods by which we define ontologies have mostly focused on the object view and neglect field data. This does not only exclude a huge amount of relevant datasets but also fails to prevent semantic interoperability problems at an early stage.

The Linked Data postulate of assigning URIs to identify entities is a good example showing how much current work is focused on objects. It is not clear how to assign URIs to field data. For instance, remote sensing instruments collect data in a range defined by their swath width and cut down the data into manageable chunks. The resulting data scenes often span thousands of square miles and consist of millions of pixels holding the measured values. Assigning URIs to each of these pixels is meaningless. The same is true for assigning a single URI to each scene as they are artifacts of data collection and often dissect relevant features, e.g., rivers.

The observations just made prompt the question, how field-based data can become a first class citizen of Linked Data and the Semantic Web, in order to transcend the current object-centric perspectives.

Accuracy of categorization

Understanding data quality is crucial for the integration and analysis of heterogeneous data; positional accuracy, attribute accuracy, logical consistency, and completeness come to mind. In terms of georeferenced data, for example, positional error distribution is measured by comparing digitized locations to those on the ground, i.e., by ground truthing to higher-precision measures or convention. Similarly, logical consistency is determined by (topological) rules; e.g., roads and buildings must not overlap. However, we lack methods to describe the *semantic accuracy* of categorical data in the same way.

To give a concrete example, if a dataset categorizes neighborhoods in a city according to a particular land cover ontology and declares a certain area as *21. Developed, Open Space* while remote sensing data shows a highly developed area with apartment complexes and industry, then the assigned category is more "off" than a second dataset classifying the same area as *23. Developed, Medium Intensity*. Note that this does not require a *true* categorization, but rather a reference dataset. Similarly as positional accuracy is measured via spatial distance, semantic similarity and analogy have been proposed as a semantic distance between classes defined in ontologies [24].

While we have only discussed the accuracy of category assignment here, the general challenge will be to develop measures for ontology quality, fitness for purpose, conceptual drift and evolution, as well as a more detailed understanding of the social construction of geographic feature types (which especially also includes events and processes).

Exploratory interfaces

The ability to semi-automatically create faceted user interfaces based on the underlying ontologies is one of the great achievements of the Semantic Web. This is made possible by shifting parts of the application logic into the data and combining it with well standardized reasoning services and querying capabilities. However, more complex queries and scientific workflows require new, exploratory interfaces, dialog systems, and new reasoning services based on analogies and similarity [24].

For example, researchers may want to evaluate a particular finding made in their study region by searching for related regions. In terms of analogy-based reasoning, they are searching for a region that differs in some properties, e.g., location, culture, mean temperature, or population density, while the properties to be evaluated, e.g., the spread of a disease, remain invariant. Semantic similarity enabled interfaces can assist users in browsing and navigating data while requiring less knowledge about the underlying ontologies. Besides improving semantics-based search interfaces, they also enable paradigms such as query-by-example. Instead of explicitly querying for particular terms or classes, users can provide concrete examples which are then used to automatically extract their commonalities, i.e., those properties that should remain invariant, and exploit them for information retrieval or recommender systems [24].

Combining analogy reasoning and similarity-based query-by-example, enables searching for the *Riviera of the United States*[2] or the *Deepwater Horizon oil spill of the 1980s*. In each of these examples, and key for the construction of analogies, particular characteristics remain invariant or are generalized to their super-classes and super-relations, while other characteristics are adjusted by the system to compute results.

Dynamic typing & the dynamic nature of links

One of the core claims of Linked Data is that by breaking up *data silos* we enable new and dynamic ways in which data can be reused and combined. A typical example is the extraction and triplification of data from Web document. However, while separating data from documents improves accessibility it puts more burden on the interpretation. Documents encapsulate information by providing reference frames and context for the inherent data and, thus, support the process of interpretation, i.e., the reconstruction of what was meant with the data. As a consequence, it is theoretically possible to run queries over the Linked Data Web that span over multiple sources to answer complex questions and establish new links between data on-the-fly, in practice, however, retrieving meaningful results is challenging or even impossible. We assume that these problems can be approached by ontologies and semantic annotations, i.e., by developing more intelligent and machine-interpretable metadata.

Surprisingly, in some cases this may have the opposite effect. If we type data too early and with classes that are loaded in terms of their ontological commitments and domain specific views, we may restrict reusability instead of fostering it. In fact, scientists should rather prefer those classes that can be deconstructed to observations. They should also publish provenance data describing the used procedures and workflows, as well as further contextual information that may assist in the interpretation of data. For instance, whether a particular area that is covered with trees constitutes a *forest* [26], should be determined when the data is reused in a given context and not prematurely declared while publishing the data.

This does not mean that ontologies should not introduce classes such as *Forest* but that these ontologies should be available as external sources in ontology repositories and combined with the data at runtime. To realize the vision of a Digital Earth as knowledge engine, scientists should be able to create ontologies for their specific needs (or reuse existing ontologies) and then integrate Linked Data based on the selected ontologies. As will be argued below, ontology design patterns and micro-ontologies may enable such a flexible selection and combination. To give a con-

[2] Which is claimed to be Santa Barbara, but this can be tested now.

crete example, loading the same forestry dataset from the Linked Data Web using a forest definition from Germany produces radially different forests than applying the Portuguese definition to the same data [26]. What can the Semantic Web learn from dynamic and late typing approaches that have been successful in software engineering? How do we determine the 20% of knowledge engineering that enables 80% of semantic interoperability without over-engineering and restricting reusability?

The same argument made in proposing to uncouple domain-specific ontologies from the data level and its observation-driven ontologies, does also hold for links. Many data providers, such as libraries or scientists, wanting to share their data do not realize that linking on the Web is a highly dynamic process. The target of their link may change in content frequently and is outside of their control – even more, relations such as `owl:sameAs` are symmetric. How can this be brought in line with the quality and archival needs of the scientific community.

Knowledge representation and reasoning

The idea of a Digital Earth as knowledge engine exposes several issues which have so far not been resolved in Knowledge Representation and Reasoning (KR) research – or in fact have even been neglected. We discuss some of them in the subsequent paragraphs.

Real data – even if we abstract from syntactic issues – is usually noisy. *Noisiness* is here used as a catch-all phrase indicating all kinds of issues which make data integration and interoperability difficult, including measurement and modeling errors, use of different design patterns, different viewpoints (i.e., semantic heterogeneity), vagueness (which includes uncertainty in the fuzzy set theory sense, and probabilistic data), and so forth. While some of these issues have been studied on the schema level (see below), KR research has not yet produced a significant body of research results which deals with *data* noisiness in the sense of ground facts, i.e., the *ABox*. To a certain extent, some approaches from fuzzy logic, probabilistic logics, and from inconsistency handling can be carried over. Some approaches related to, e.g., default (and related) logics may be helpful for bridging semantic heterogeneity and diverging design patterns. But overall, there is little work or experience in handling data noisiness at large scale, and in uncontrolled settings like the Digital Earth. It could be conjectured that the reason for this neglect lies in the fact that these issues just have not really arisen in practice so far. However, the vision presented here – and in more generality problems related to the handling of Big Data or Linked Data – do raise these issues and give the finding of solutions to them immediate practical relevance.

If we move from the data level to the schema level (*TBox*), i.e., to domain modeling with ontologies, then again we find that some issues important for the realization of a Big Data knowledge engine are underdeveloped, i.e., the state of the art does not provide ready-to-use solutions for some central problems. Examples for such central issues can easily be found when considering some aspects of human cognition which would have to be reflected in KR solutions. In particular, humans excel in navigating different viewpoints, different scales, and different contexts, which are all aspects which could be summarized under the notion of semantic heterogeneity. We humans seem to be able to effortlessly integrate such semantically heterogeneous information in most situations. Somewhat more tangible based on the current state of the art appear issues like the handling of stereotypes or defaults, the mixing of open- and closed-world perspectives, and dealing with vagueness. However, even with respect to the latter list, it is rather unclear how to adapt the state of the art to a modeling context with deliberate semantic heterogeneity – and scalability and usability issues also remain to be resolved.

The systematic use of well-designed ontology design patterns may provide a partial technical solution to dealing with knowledge integration and data interoperability in the presence of semantic heterogeneity. Indeed, ontology design patterns which have been created based on a consensus by different stakeholders are naturally amenable to different viewpoints, yet provide a single pattern across usages and domains which can be leveraged for integration. This seems to indicate that they are more suitable for heterogeneity preservation when integrating knowledge, than the use of foundational ontologies, which necessarily forces the strong ontological commitments made for the foundational ontology onto the domain ontologies. A hope would be, that a critical supply of (application domain specific) ontology design patterns could give rise to a network of local micro-ontologies which capture specific definitions and ontological commitments required for a specific modeling or engineering task, in such a way that these micro-ontologies are horizontally interconnected and interconnectible through the fact that they are based on the same design patterns [22].

Indeed a systematic use of ontology design patterns would be a much preferable alternative to the common *shallow modeling first, deep modeling (hopefully) later* approach, which is bound to create more trouble than solutions [20]. The trouble comes from the experience that it is usually impossible to start modeling with an inexpressive language, in the hope to be able to add stronger ontological commitments later: after initial "shallow" modeling it will usually turn out that due to the initial ambiguity of terms and modeling patterns, the resulting knowledge base is no longer semantically homogeneous, and thus cannot be semantically strengthened in a way which is consistent with the knowledge and data already present. A recent and rather prominent example for the fallacies in this approach is the use of the `owl:sameAs` language construct in Linked Data: While it occurs in very substantial quantities,[3] its usage is mostly informal and in particular is not aligned with the formal semantics that it should inherit from OWL [14] – a problem which, at hindsight, could have been avoided by taking deep semantics, in this case, the formal meaning which `owl:sameAs` has been given by the OWL formal semantics [27], into consideration in the first place. By adhering to well-designed ontology design patterns, modeling could at first be restricted to dealing with the patterns, while well-designed patterns will easily be amenable to semantic strengthening.

On the algorithmic side, it has been proposed to transcend the deductive paradigm by viewing ontology reasoning, at least partially, from an information retrieval perspective [18]. The key idea is to understand a deductive reasoning task (which in its basic form yields a yes or a no as answer) as a classification problem (classify the input query as "yes" or as "no"). From this perspective, deductive reasoning can at least in principle be approached with information retrieval (i.e., non-deductive) methods, the performance of which can be assessed in terms of precision and recall, with the output of a deductive reasoner as baseline. It could then be hoped, that such non-deductive reasoning approaches could carry over to noisy or semantically heterogeneous settings. Indeed, the potential power of such non-deductive methods for question answering has been shown at scale, and in an impressive way, by the performance of IBM's Watson system in the Jeopardy! game show.

Ontology alignment

Ontology alignment [9] refers to the creation of formal relationships between entities in different ontologies. In the simplest and most well-studied case, these relationships take the form of subsumption (`rdfs:subClassOf`), class equivalence (`owl:equivalentClass`), or equality (`owl:sameAs`). Even for such simple relationsships, which are well-studied in the ontology alignment literature, the noisy nature of Linked Data initially prevented many established systems from performing well, so that new approaches had to be established [19,21].

In order to deal with semantic heterogeneity, ontology design patterns, and micro-ontologies, the simple ontology alignment setting just described needs to be lifted considerably towards more complex alignments. Complexity, in this case, refers to at least the following two dimensions.

(i) On the one hand, alignments need to be able to map different modeling choices onto each other, by making use of complex logical expressions. E.g., an example in [18] shows that the seemingly simple piece of information that "Nancy Pelosi voted in favor of the 2009 health care bill" is modeled in GovTrack using 8 rather convoluted RDF triples, while a straightforward attempt to model the same piece of knowledge, at the presented granularity, would probably need much fewer triples with a much clearer structure. A complex alignment (expressible in OWL [16] or RIF [5]) could then easily be described which maps one structural representation to the other and vice versa. While this example resides on the data level, it easily generalizes to the schema level, where such complex alignments promise to be even more powerful.

(ii) On the other hand, alignment primitives need to be established which differ in semantic strength, such that they can be used to effectively map between ontologies which are semantically heterogeneous. An early but limited example of this is C-OWL [6], which can be used to align ontologies in such a way that the combined knowledge remains (somewhat) usable even if some of the ontologies in the combination are inconsistent. In a similar, but more fine-grained way, ontology alignment methods need to be established which control not only potential causes for inconsistencies, but also cater for *default* alignments (which may have exceptions), stereotypes, etc.

The research issue of providing such kinds of alignment primitives can in fact not be separated from research into dealing with micro-ontologies – essen-

[3] See http://stats.lod2.eu/.

tially, the same primitives which will be useful for (weakly; semantically heterogeneously) integrating micro-ontologies will also be the primitives which have to be studied for ontology alignment. As before, the role of well-designed ontology design patterns as kernels for integration and alignment should not be underestimated; indeed in the case of reused patterns, some part of the alignment problem becomes almost trivial.

The big research issue with both kinds of complex ontology alignments is, obviously, how to create such alignments using automated methods. Indeed, there is, as yet, embarrassingly little research on this issue.

Conclusion

Realizing the laid out vision of a Digital Earth as a knowledge engine requires to develop generic methods and tools driven by a concrete but vast goal. For example, with respect to ontology reasoning, this requires the development of practically applicable integrated methods for dealing with stereotypes and defaults, with weak notions of equivalence, with noise and inconsistency in the data, etc., all of which have been studied in the ivory tower but have not yet had substantial impact on practice [18]. Similar advances are required in other fields, *driven by a concrete application vision*. Such a channeling of resources has the potential to be catalytic for future research and to be a showcase for the added value and strength of the Semantic Web, in a similar way in which RoboCup transformed robotics.

If we resist the temptation to follow the seemingly simple path by trying to *resolve* semantic heterogeneity, but instead accept heterogeneity as the motor of science, we can expect that work on semantics-driven integrity constraint checking, ontology matching and alignment, reasoning in the presence of inconsistencies and uncertainty, defaults, semantic negotiation, similarity and analogy reasoning, bottom-up semantics, inductive approaches, and so forth will play a key role in interdisciplinary research. It is a common misconception that interoperability could only be achieved by a rigid standardization process that results in a small number of foundational and domain level ontologies. Instead, we should exploit the power of Semantic Web technologies and knowledge patterns to directly establish interoperability between purpose-driven ontologies without having to agree on a universal level before.

Finally, as a research community, we need to emphasize the paradigm shift proposed by the Semantic Web and Linked Data and abstract from specific technological solutions. We need to explain how to derive ontologies from scientific workflows and data and demonstrate the added value of publishing Linked Data in a way that relates to the immediate needs of individual researchers. What are the minimal requirements for these researchers to actively participate and contribute their data?

Acknowledgements

Parts of this document have been drafted while the authors met at Dagstuhl Seminar 12221 on *Cognitive Approaches for the Semantic Web*. 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 authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] C. Anderson and M. Wolff, The web is dead. Long live the internet, *Wired Magazine* (September 2010).

[2] L.W. Barsalou, Grounded cognition, *Annual Review of Psychology* **59**(1) (2008), 617–645.

[3] T. Berners-Lee, J. Hendler, and O. Lassila, The Semantic Web, *Scientific American* **284**(5), (May 2001), 34–43.

[4] C. Bizer, T. Heath, and T. Berners-Lee, Linked Data – The story so far, *International Journal on Semantic Web and Information Systems* **5**(3) (2009), 1–22.

[5] H. Boley, G. Hallmark, M. Kifer, A. Paschke, A. Polleres, and D. Reynolds, eds, *RIF Core Dialect*, W3C Recommendation 22 June 2010, 2010. Available from http://www.w3.org/TR/rif-core/.

[6] P. Bouquet, F. Giunchiglia, F. van Harmelen, L. Serafini, and H. Stuckenschmidt, Contextualizing ontologies, *Journal of Web Semantics* **1**(4) (2004), 325–343.

[7] B. Brodaric and M. Gahegan, Ontology use for semantic e-science, *Semantic Web* **1**(1–2) (Apr. 2010), 149–153.

[8] M. Craglia, M. Goodchild, A. Annoni, G. Camara, M. Gould, W. Kuhn, D. Mark, I. Masser, D. Maguire, S. Liang, and E. Parsons, Next-generation digital earth, *Int. J. Spatial Data Infrastructures Research* **3** (2008), 146–167.

[9] J. Euzenat and P. Shvaiko, *Ontology Matching*, Springer-Verlag, Heidelberg (DE), 2007.

[10] D. Ferrucci et al., Building Watson: An overview of the DeepQA project, *AI Magazine* **31**(3) (2010).

[11] M. Goodchild, Citizens as sensors: The world of volunteered geography, *GeoJournal* **69**(4) (2007), 211–221.

[12] M. Goodchild, M. Yuan, and T. Cova, Towards a general theory of geographic representation in gis, *International Journal of Geographical Information Science* **21**(3) (2007), 239–260.

[13] A. Gore, The digital earth: Understanding our planet in the 21st century, 1998.

[14] H. Halpin, P.J. Hayes, J.P. McCusker, D.L. McGuinness, and H.S. Thompson, When owl:sameas isn't the same: An analysis of identity in Linked Data, in: *The Semantic Web – ISWC 2010 – 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7–11, 2010, Revised Selected Papers, Part I*, P.F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J.Z. Pan, I. Horrocks, and B. Glimm, eds, Lecture Notes in Computer Science, Vol. 6496, Springer, Heidelberg, 2010, pp. 305–320.

[15] J. Hawkins and S. Blakeslee, *On Intelligence*, Times Books, 2004.

[16] P. Hitzler, M. Krötzsch, B. Parsia, P.F. Patel-Schneider, and S. Rudolph, eds, *OWL 2 Web Ontology Language: Primer*, W3C Recommendation 27 October 2009, 2009. Available from http://www.w3.org/TR/owl2-primer/.

[17] P. Hitzler, M. Krötzsch, and S. Rudolph, *Foundations of Semantic Web Technologies*, Chapman & Hall/CRC, 2009.

[18] P. Hitzler and F. van Harmelen, A reasonable Semantic Web, *Semantic Web* **1**(1–2) (2010), 39–44.

[19] P. Jain, P. Hitzler, A.P. Sheth, K. Verma, and P.Z. Yeh, Ontology alignment for linked open data, in: *The Semantic Web – ISWC 2010 – 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7–11, 2010, Revised Selected Papers, Part I*, P.F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J.Z. Pan, I. Horrocks, and B. Glimm, eds, Lecture Notes in Computer Science, Vol. 6496, Springer, Heidelberg, 2010, pp. 402–417.

[20] P. Jain, P. Hitzler, P.Z. Yeh, K. Verma, and A.P. Sheth, Linked Data is Merely More Data, in: *AAAI Spring Symposium 'Linked Data Meets Artificial Intelligence'*, AAAI Press, 2010, pp. 82–86.

[21] P. Jain, P.Z. Yeh, K. Verma, R.G. Vasquez, M. Damova, P. Hitzler, and A.P. Sheth, Contextual ontology alignment of LOD with an upper ontology: A case study with Proton, in: *The Semantic Web: Research and Applications – 8th Extended Semantic Web Conference, ESWC 2011, Heraklion, Crete, Greece, May 29–June 2, 2011, Proceedings, Part I*, G. Antoniou, M. Grobelnik, E.P.B. Simperl, B. Parsia, D. Plexousakis, P.D. Leenheer, and J.Z. Pan, eds, Lecture Notes in Computer Science, Vol. 6643, Springer, Heidelberg, 2011, pp. 80–92.

[22] K. Janowicz, The role of space and time for knowledge organization on the Semantic Web, *Semantic Web* **1**(1–2) (2010), 25–32.

[23] K. Janowicz, Observation-driven geo-ontology engineering, *Transactions in GIS* **16**(3) (2012), 351–374.

[24] K. Janowicz, M. Raubal, and W. Kuhn, The semantics of similarity in geographic information retrieval, *Journal of Spatial Information Science* **2** (2011), 29–57.

[25] W. Kuhn, Geospatial semantics: Why, of what, and how? in: *Journal on Data Semantics III*, S. Spaccapietra and E. Zimányi, eds, Lecture Notes in Computer Science, Vol. 3534, Springer, Berlin/Heidelberg, 2005, pp. 587–587.

[26] G. Lund, Definitions of forest, deforestation, afforestation, and reforestation. [online] gainesville, va: Forest information services. Available from the world wide web: http://home.comcast.net/ gyde/DEFpaper.htm, Technical report, 2012.

[27] B. Motik, P.F. Patel-Schneider, and B. Cuenca Grau, eds, *OWL 2 Web Ontology Language: Direct Semantics*, W3C Recommendation, 27 October 2009. Available at http://www.w3.org/TR/owl2-direct-semantics/.

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

[29] S. Scheider, Grounding geographic information in perceptual operations. PhD thesis, University of Münster, Germany, Technical report, 2011.

[30] C. Schlieder, Digital heritage: Semantic challenges of long-term preservation, *Semantic Web* **1**(1–2) (2010), 143–147.

[31] E. von Glasersfeld, Reconstructing the concept of knowledge, *Archives de Psychologie* **53**(204) (1985), 91–101.