Linked Data - A Paradigm Shift for Geographic Information Science

Werner Kuhn[1,2], Tomi Kauppinen[3,4], and Krzysztof Janowicz[2]

[1] Center for Spatial Studies, University of California, Santa Barbara CA, USA
[2] Department of Geography, University of California, Santa Barbara CA, USA
werner|janowicz@ucsb.edu
[3] Cognitive Systems Group, Universität Bremen, Germany
[4] Department of Media Technology, Aalto University School of Science, Finland,
tomi.kauppinen@aalto.fi

Abstract. The Linked Data paradigm has made significant inroads into research and practice around spatial information and it is time to reflect on what this means for GIScience. Technically, Linked Data is just data in the simplest possible data model (that of triples), allowing for linking records or data sets anywhere across the web using controlled semantics. Conceptually, Linked Data offers radically new ways of thinking about, structuring, publishing, discovering, accessing, and integrating data. It is of particular novelty and value to the producers and users of geographic data, as these are commonly thought to require more complex data models. The paper explains the main innovations brought about by Linked Data and demonstrates them with examples. It concludes that many longstanding problems in GIScience have become approachable in novel ways, while new and more specific research challenges emerge.

Keywords: Linked Spatiotemporal Data, Resource Description Framework, Geo-Semantics

1 Introduction

Linked Data is an example of a technological innovation that transforms the way we think about information and its role in society, in particular geographic information. However, discussions on it tend to focus on technical aspects, such as how to convert existing data sets or how to deal with semantics in shared vocabularies and ontologies. This paper explains why the difference between traditional data holdings and Linked Data repositories is more than one of formats and is largest for sophisticated types of information, in particular information with spatial and temporal components.

With the adoption of Linked Data, the familiar complexities of conceptual database schemata for spatial data can safely remain internal to organizations, from where they have been too hard to share anyway. Their externally relevant contents get streamlined into the open and more manageable form of vocabulary definitions. Users of Linked Data do not need to be aware of complex schema information to use data adequately, but ”only” of the semantics of types and predicates (such as `isLocatedIn`) occurring in the data. While many questions remain to be answered about how to produce and maintain vocabulary specifications [10], the elaborate layering of syntactic, schematic, and semantic interoperability issues [5] has simplified to a single common syntax (RDF), the irrelevance of traditional schema information outside a database, and a focus on specifying and sharing vocabularies.

This simplification is more dramatic for spatially and temporally referenced data (with their complexities in the form of geometries and scale hierarchies) than it is for, say, financial data. The resulting paradigm shift, from distributed complex databases accessed through web services that expose schemata to knowledge represented as graphs, whose links can be given well-defined meaning, radically changes some of the long-standing problems of GIScience and GIS practice. This paper attempts to raise the level of discussion on Linked Data from the ”how” to the ”why” by describing the changes in perspective on some deeper issues of GIScience. It also summarises new problems and research questions arising from the paradigm shift.

The discussion should be seen against the broader background of commonly identified limitations of existing data models for spatial and spatio-temporal data, including that

- access to data is software-dependent
- metadata and schemata are separate entities
- data models mix concerns of semantics and data management
- the semantics of terms remains implicit or hard to share and reason with
- data are seen as provider-independent truths, though they often contradict
- there are too few simultaneously accessible viewpoints or versions of data
- global, unique identifiers are hard to obtain and not encouraged
- valuable data sets remain isolated and hard to integrate
- the emphasis on consistency and quality restricts data availability
- data about a particular topic, place, or period are hard to find
- incentives for producing metadata and enabling reuse are lacking.

.

Linked Data is not a pandora box removing these obstacles to producing, accessing, and using spatial data. But it is more than a set of new technologies, as it substantially changes how we deal with these problems. For example, geographic information has long been recognized as a powerful ”glue” to integrate information across domains, but putting this vision in place was often too hard. With Linked Data, the gluing function of spatial and temporal referencing has finally become reality [14], as the Linked Data Cloud[5] allows for linking any data to geographically referenced data, such as linked geodata[6].

After introducing the basics of Linked Data in the next section, sections three to nine discuss the impact of Linked Data on how we understand spatial and

[5] http://datahub.io/group/lodcloud

[6] http://linkedgeodata.org/

spatio-temporal information and the issues surrounding it: provenance, consistency, metadata, semantics, maintenance, data publishing, and data integration. We conclude with some new or revised research challenges and a summary of the points made.

2 Linked Data in a Nutshell

Linked Data is the name for a collection of design principles and technologies centered around a novel paradigm to publish, retrieve, reuse, and integrate data on the Web. In contrast to the Document Web, the Web of Linked Data aims at establishing named and directed links between typed data. For example, a normal Web page about Portsmouth (such as `http://en.wikipedia.org/wiki/Portsmouth`) may link to another page about Hampshire (such as `http://en.wikipedia.org/wiki/Hampshire`). For a machine, the intended meaning of such links is difficult to interpret and the Web pages can only be consumed as integral units of text or other media. On the Linked Data Web, by contrast, the link between Portsmouth and Hampshire would be directed and labeled, for example, forming the statement that Portsmouth is located in Hampshire. Additionally, the two places would be typed, e.g., as city and county, jointly leading to the statement that the city of Portsmouth is located in the county of Hampshire. Finally, the predicate `isLocatedIn` could be defined as a transitive relation in an ontology. Thus, in conjunction with a statement that Hampshire county is located in the UK, one could automatically derive the new statement that Portsmouth is located in the UK.

Given that three elements constitute each piece of information in Linked Data, one refers to such statements as *triples*, consisting of a subject (Portsmouth), a predicate (`isLocatedIn`), and an object (Hampshire). This syntax, which happens to be the simplest form in which statements can be made in natural language, has thus been carried over to the world of data. The data model for triples is the so-called Resource Description Framework (RDF).

Tim Berners-Lee established four principles of Linked Data [3]:

- Uniform Resource Identifiers (URI's) should be used to denote things.
- HTTP URI's should be used so that these things can be referred to and dereferenced (looked up) by human users and software agents.
- W3C standards such as RDF or OWL should be used to provide information about the things when their URI's are dereferenced.
- Data about anything should link out to other data, using their URI's to create a densely interconnected graph of knowledge (the so-called Linked Data Cloud).

As these principles are expressed in the technical jargon of the Web, it helps to relate them to spatial data and entities in geographic space. According to the principles, an entity in the physical world, such as the historic ship HMS Victory, should be identified by a globally unique URI. As the HMS Victory is not an information resource, one cannot directly retrieve information about it using a browser or Linked Data tools. However, when visiting the ship's URI, the responding Web server can redirect (HTTP code 303) the visitor to an information resource, such as an RDF document containing statements about HMS Victory or an HTML page that renders RDF in a human readable way[7]. (As a URI for HMS Victory, we use `http://dbpedia.org/page/HMS_Victory`, which can be abbreviated to `dbpedia:HMS_Victory`, thanks to the predefined name space dbpedia[8]).

Linked Data can be queried using SPARQL[9], a query language for RDF. GeoSPARQL adds the possibility to query over topological relations and thus enriches SPARQL by quantitative reasoning. So far, GeoSPARQL support is limited but some reference implementations, such as Parliament, have been proposed and implemented as free and open source [2]. The following SPARQL query example retrieves all predicates and objects of statements that have dbpedia:HMS_Victory as a subject.

```
SELECT * WHERE {
  dbpedia:HMS_Victory ?predicate ?object
}
```

The queried statements might contain information about the ship, e,g., when it was laid down or the battles it participated in. While the first case can be represented by a single date, e.g., using XSD date type, the linked historic battles could themselves be represented by URI's[10], linking to actors involved in the battles, such as Vice-Admiral Horatio Nelson. A triple may state that the HMS Victory is located at Portsmouth, UK, which in turn leads to more resources about that city, its population, and so forth, contributing to the more and more densely interconnected Linked Data Cloud.

Linked Data is usually stored in so-called triple stores and accessed via so-called SPARQL endpoints. The ontologies that allow human users and machines to understand which concepts and predicates can be queried, and how they are formally defined, are described using languages such as the Web Ontology Language (OWL).

Listing 1.1 shows five RDF triples in Turtle syntax[11]. Some of them are assertions, e.g., that Horatio Nelson died in the Battle of Trafalgar, while others are taxonomic, e,g., that naval battles are special battles. One can derive new statements from those triples; for instance, one can automatically infer that Nelson died in a battle.

[7] See `http://live.dbpedia.org/page/HMS_Victory`

[8] `http://dbpedia.org/sparql?nsdecl`

[9] DBPedia has an open SPARQL endpoint at `http://live.dbpedia.org/sparql`

[10] though DBPedia represents the battles only as literal values, not enabling queries to follow links

[11] `http://en.wikipedia.org/wiki/Turtle_(syntax)`

```
ex:HoratioNelson ex:diedIn ex:BattleOfTrafalgar .
ex:BattleOfTrafalgar ex:during ex:NapoleonicWars ;
                     rdf:type ex:NavalBattle .
ex:NavalBattle rdfs:subClassOf ex:Battle .
ex:diedIn rdfs:subPropertyOf ex:participatedIn .
```

Listing 1.1. Example of RDF statements.

The discussion in this paper focuses on data exposed as Linked Data, regardless of a possible co-existence of the same information in other data models (e.g., in GIS, databases, or file systems) and regardless of whether the data are considered open or not. Linked Data is in fact most typically treated as (but not limited to) a secondary, openly available, exposure of contents that are also held in other formats, typically enabled through government subsidies or open source communities.

3 Data are Statements

Linked Data express claims [14], i.e. statements made by somebody somewhere at some time. Meta-information about this context is obtainable through the owner of the URI at which the statements reside and the place and time of their publication. This idea holds the key for solving provenance, currency, and versioning problems. Provenance or lineage (the origin and chronology of data) remains a thorny issue in theory and practice around geographic data and data in general. Linked Data deals well with provenance, but remains somewhat weak in dealing with the temporal evolution of knowledge.

For an example of how to obtain provenance information, consider the triple stating that Portsmouth is located in Hampshire. The author of such a statement is typically made explicit only at the level of whole data sets. The statement about Portsmouth could be (but actually is not) part of Ordnance Survey's publicly available linked geodata[12]. According to http://data.ordnancesurvey.co.uk/datasets/os-linked-data, Ordnance Survey did make the taxonomic statement, however, that the City of Portsmouth is a Borough, on October 25th 2010, and confirmed it on May 10th, 2013.

The nature of triples as statements, as obvious as it is from their subject-predicate-object syntax, tends to be mixed up with the conventional view of data as facts, even in the specialized literature discussing how to reason with Linked Data. Triples are still too often seen as single, objective, eternal truths about some contents, irrespective of authorship or date or other contextual and quality aspects. The implicit assumption is that they will conveniently be overwritten or forgotten if better data become available.

When triples are instead seen as expressions of beliefs held by individuals or organizations at some point in time, the Linked Data paradigm easily admits reviewing, commenting, revising, and extending information. If Ordnance Survey, for example, makes statements about the geography of the UK, these come

[12] http://data.ordnancesurvey.co.uk/

with a significant level of authority and trustworthiness. But they still express a view of the world held by a (professional, authoritative) organization at some point in time, subject to disagreement, revision, and improvement.

Trust and reputation are naturally attached to the authors of data, propagated to their statements, and calibrated through links and their weights in search engines. The Linked Data paradigm is ideal to express and reason with such information. There are many solutions for handling and increasing trust: 1) professional and volunteered curation of data, 2) inconsistency checks via automated reasoning, 3) applications and services using the data. These processes can also take the form of syntactic checks. Similarly, reputation should increase based on communication, i.e. sharing evidence about the level of trust: for instance communicating that the data in some source is curated.

Thanks to the Linked Data paradigm, digital maps and other geospatial models can now be seen as sets of statements made by authors with some reputation at some well-defined points in time [25]. Ideally, these statements will never be removed, because it is valuable (and sometimes legally required or otherwise essential) to know about previous world views or states of affairs. New insights may be gained, the world or the ways to describe it may change, and statements can simply be added to the Linked Data Cloud.

From a technical point of view, with RDF 1.1, the possibility to group statements in data sets attributed to a well-defined (by URI) provenance has now become a standard solution (using multiple graphs)[13].

4 Statements can Contradict

Another long standing concern in GIScience is that of the consistency of its models. Following Allemang and Hendler's AAA slogan that '**A**nyone can say **A**nything about **A**ny topic' [1], consistency is not a key concern for Web-scale systems. Given the variety of sources, perspectives, granularities, and data creation and curation procedures, the collection of statements in the Web can, do, and will contain contradictions. Early on, the Semantic Web community decided to handle this challenge not by stricter models and model checking but by adopting the Open World Assumption (where the absence of a statement does not imply its falsehood). It focuses on inferential semantics, deferring consistency checking to the level of ontology engineering.

To give a concrete example for the consequences of this decision, consider the death of a prominent person. This information may be updated at Wikipedia (and thus DBpedia live[14]) within minutes but may take longer to be integrated into other data repositories. News stations may decide to wait for independent confirmation of the event. Thus, the global knowledge graph will contain information that the given person is alive and dead at the same time. A Web-scale system such as the global graph of Linked Data still has to be able to function despite such apparent inconsistencies.

[13] https://dvcs.w3.org/hg/rdf/raw-file/default/rdf-dataset/index.html

[14] http://live.dbpedia.org

On the terminological level, the Open World Assumption ensures that a lack of knowledge or the temporal unavailability of a certain data repository does not imply that a statement is false. In contrast, under the Closed World Assumption statements that are not known to be true, are assumed to be false. In the example, the absence of a date of death does not imply that the person is still alive, just that it is unknown whether the person is alive or dead (until a statement clarifies this either way).

In addition to such temporal aspects, space and place also play a key role in interpreting statements. For instance, news agencies from two different countries may have diverging opinions (on a cause of death or anything else). These regional (as well as temporal) differences can go as far as authorities disagreeing at the terminological level. The case of ”freedom fighters” versus ”terrorists” used to label the same people over time is an infamous example. Two data providers at the same time, or the same provider at two different times, may *classify* a particular individual in different ways. This leads to an extended version of the AAA slogan to an AAAAA view of '**A**nyone can say **A**nything about **A**ny topic at **A**ny time and **A**nywhere' [15]. The lesson from such examples is that spatial and temporal indexing of statements is important for reasoning, even where the topics reported do not seem to be spatial or temporal.

Additionally, and in clear contrast to previous knowledge engineering paradigms, the Linked Data community has taken the stance that there is no need for linking to abstract top-level or domain-level ontologies (while retaining the benefits of designing ontologies along well-defined upper level distinctions). It has invested instead into research on ontology alignment, data-driven knowledge patterns, and query federation. Consequently, what appears to be the same statement in two triple stores can have vastly different interpretations. Consider, for instance, a triple stating that Horatio Nelson died on the deck of the HMS Victory, and a second triple that the HMS Victory is located at Portsmouth, UK. Depending on the choice of ontology axioms used to interpret these statements, including decisions on how to assess the temporal validity of statements, one can infer that Nelson's place of death is in the UK or not. Whether this is an unintended logical consequence is left to decide at the terminological level, i.e., by considering and adapting the ontologies used. This attitude fosters reusability and integration. Its downside is that the (mis)use of so-called co-reference resolution (e.g., via `owl:sameAs`) may hamper conflation; whether two resources actually refer to the same entity is often a very complex decision to make; see Haalpin et al. [13] for a detailed discussion.

5 Metadata are Data

GIScience and GIS practice distinguish data (such as on geographic features and their attributes) from metadata (such as on the creation dates and names of creators of the data). The two types of data tend to be captured and kept in separate locations, models, formats, granularities, and business models. Linked Data blurs this largely artificial distinction, as each statement can be semantically typed and annotated. It is still possible to create provenance information about collections of statements (RDF documents), or SPARQL endpoints serving certain data. This can, for example, be done using the Prov ontology [22]. Provenance statements, however, are statements of the same nature and form as all other statements in RDF (which itself evolved from a metadata format). It is possible to follow links from data to metadata and back. It is also possible to define what is considered metadata in some application scenario and what is considered data. Finally, on the terminological level, ontologies developed in OWL or the RDF Schema language (RDFS) can also be queried to retrieve data about individuals as well as about concepts (which would be considered metadata) and one can pose queries that filter the results by criteria about concepts and data.

Metadata about statements (spatial or not) are themselves often spatial and temporal: when did the information become known? when and where did an event occur? what was its duration? what else happened during that time or at that place? These spatial and temporal aspects of metadata make spatio-temporal computing attractive for information in general, beyond properly geographic data. They can be better exploited now that the data about them is part of content data. Through the tight integration of data and metadata in a single and simple format (RDF), we are now in fact applying GIScience methods to information infrastructures in general [15].

6 Semantics is in Predicates, not Schemata

One of the biggest innovations of Linked Data is the way semantics gets handled. Traditional geographic data modeling wisdom (to be found in any GIScience textbook or course) has it that a conceptual database schema, together with a data dictionary, is the best way to capture what is meant by the data. Ontologies may be used to expose semantics and allow for machine reasoning about it, but they are typically not seen as having reached the expressive power or practicality of database schemata and dictionaries. A database schema, in this sense, describes the structure of data, while an ontology provides a specification of the intended meaning of terms. Thus, the triple structure is a schema, while ontologies (expressed in OWL or other languages) specify the types and predicates used in triples.

One problem with the database schema approach to semantics is that data often leave their native environments and get repackaged in forms which may or may not capture the intended semantics adequately. Supplying them with some schema information (say, in XML-based form) is normally not enough. Database schemata, even in their layered form standardized by ANSI and ISO (conceptual - logical - physical), fail to separate the concerns of data organization from those of semantics. They do a great job on the former, but a notoriously poor one on the latter task, especially when data leave their native environments. Semantics has very little to do with how data are structured, and much more with how terms are being used in the data (and their structuring).

Linked Data provides this missing semantic link for spatial (as well as any other types of) data by connecting statements to definitions of the predicates used in the triples. Since their conceptual schema is the simplest possible one, that of a triple with a subject, predicate and object, nothing else needs to be stated about structure or schemata and one can concentrate all semantic efforts on capturing the intended meaning of the terms used in these three elements. Semantics that was traditionally captured in schema form (say, about cardinalities of relations) can be restated in shared vocabularies. In this form, it will be explicit and accessible to inspection and revision anywhere within and outside the organization producing or holding the data. As syntactic interoperability is largely handled by common standards such as RDF, we can entirely focus on addressing semantic interoperability.

7 Maintenance without Deletion

GIScience still relies massively on static world views—or at least on an implicit assumption that changes can be recorded through sequences of snap shots. For instance, changes related to weather (say, temperature or amount of rain) or other environmental conditions (sea level or river width) are monitored and recorded as time series of attribute values. The underlying assumption is that the essence of world knowledge is static: values of temperature change but the temperature concept itself remains unchanged and only current values of such concepts need to be stored. Many counter examples show the need for modeling changes in more sophisticated ways:

- natural and administrative regions change (they split, merge or otherwise change their form) due to human decisions or natural processes;
- connections between things change (e.g., a person moves from one affiliation to another)
- concepts change over time or acquire multiple senses (like the famous example by Frege about the concepts of *evening star* and *morning star*)

There are many such changes in the world [27] and in our way of talking about it. When formalized, they can be used for inferences. An example of this is to study [9] what topological inferences can be made when a region is cut into pieces. Linked Data allows for storing and sharing both the data used as input and the inference results. This way, inference results become links in RDF and provide a way to traverse longer paths, for instance from contemporary place names to historic ones [18].

Ideally, with Linked Data as statements, monotonicity is regained. A true statement never needs to be retracted, but the time span in which it is considered true needs to be made explicit. For place-based information, this means to understand and state the beginnings and ends of the validity of place names. The time span for the resource representing the place can be defined, for instance, as a the time when the borders of an administrative place remained unchanged

[18]. This way borders can be linked to the right place, and spatial relations—like overlap with historic or contemporary regions—can be traced over time.

New challenges emerge for this approach, such as how to deal with imprecision and uncertainty—for instance how to infer when a certain statement is valid if the beginning/end of its validity is not known exactly. A Linked Data solution for this problem is to define fuzzy temporal intervals. Instances of this concept can then be annotated with the fuzzy begin, begin, end, and fuzzy end predicates of the interval. This allows for computing with the validity of statements even if the validity itself is imprecise. Maintenance also calls for documented provenance: about who has created a statement, when it was made, or when it became obsolete [29].

8 Data Publishing and Sharing by URI

The principles of Linked Data are built around URI's. Things should be named with URI's, preferably with HTTP URI's. Accessing data by URI allows for individual statement retrieval, in contrast to the need for always downloading complete, often large datasets. SPARQL can be used to query just that part of the data that one needs for a given task, down to a single piece of information (such as what county contains Portsmouth).

It is notable in this context that SPARQL endpoints are themselves identifiable via URI's. This allows for automatic querying of data provided by endpoints, and gathering of large-scale documentation of available data of different types. However, it calls for research on representation mechanisms for spatial accuracy, resolution, and other data quality aspects. Linked data thus provides a transparent way for building future Spatial Data Infrastructures (SDI) [8].

The availability of data about all parts of a statement (i.e. subject, predicate and object) differs considerably from traditional SDI, where predicates are not described in a machine-understandable fashion. For instance, the following excerpt describes the predicate `DEFOR_2008`[15] used to describe ”new deforestation in 2008” in the Linked Brazilian Amazon Rainforest Data [17]:

```
@prefix amazon:<http://spatial.linkedscience.org/context/amazon/> .

amazon:DEFOR_2008
      rdfs:label "Percentage of new deforestation in 2008";
      amazon:aggregation amazon:Pixel ;
      amazon:columnnumber amazon:c10 ;
      amazon:source amazon:INPE ;
      amazon:timeperiod amazon:year2008 ;
      amazon:unit amazon:percent ;
      amazon:variabletype amazon:LandUse .
```

[15] Full example is accessible as Linked Data at http://spatial.linkedscience.org/context/amazon/DEFOR_2008

Deferencing by URI provides a way to check what statements are currently served by that URI. By accessing the URI of a 25kmx25km grid cell about the Brazilian Amazon Rainforest one gets the following kinds of statements[16] —i.e. aggregated information about that cell.

```
amazon:AMZ_LINKED_25K_1000
     rdfs:label "Cell 1000";
     amazon:DEFOR_2002 "0.039"^^xsd:double ;
     amazon:DEFOR_2003 "0.0030"^^xsd:double ;
     amazon:DEFOR_2004 "0.031"^^xsd:double ;
     amazon:DEFOR_2005 "0.042"^^xsd:double ;
     amazon:DEFOR_2006 "0.0050"^^xsd:double ;
     amazon:DEFOR_2007 "0.012"^^xsd:double ;
     amazon:DEFOR_2008 "0.0040"^^xsd:double .
```

Another example of Linked Data publishing is `spatial@linkedscience` [20] which contains Linked Data about papers published in the GIScience, COSIT, ACM GIS (SIGSPATIAL), and AGILE conference series. Each paper, author, and affiliation is assigned a URI. Accessing the data is again done via URI[17] to retrieve an RDF version of the data.

Communities ranging from libraries to environmental scientists have been seeking a way to identify the information resources they are dealing with. The solutions—such as Digital Object Identifiers (DOIs) for identifying outlets and papers—call for establishing registries to maintain identifiers and their mutual mappings. Linked Data facilitates such registries and the trust in them by tracing the hubs of data, similarly to search engines using HTML pages to find trusted hubs of information.

9 Data Integration by Linking

Linking enriches not only source data (that links to destination data), but also the destination data. For instance, referencing digital cultural heritage data to places creates rich descriptions of the places themselves. This allows for studying, for example, connections between places and the culture of a region.

URI's enable integration of different kinds of data not only online, but also locally. Sensitive data can make use of openly available Linked Data by sharing its URI's. This supports providing of the context (e.g., via spatial or temporal references) for the sensitive data in question, while the sensitive data itself can remain private.

[16] Full example is available at `http://spatial.linkedscience.org/context/amazon/AMZ_LINKED_25K_1000`

[17] For instance `http://spatial.linkedscience.org/context/acmgis/paper/doi10.1145/1653771.1653787`

Federated queries allow for accessing data from different SPARQL endpoints, i.e. to combine results from multiple sources. For instance the following statement documents that a grid cell is partially overlapping a municipality.

```
amazon:AMZ_LINKED_25K_1000
 tisc:partiallyOverlapping
  amazon:BRAZIL_MUNICIPALITY_1508407 .
```

By further requesting statements about the municipality[18] `amazon:BRAZIL_MUNICIPALITY_1508407` one can retrieve not only its name (Xinguara), but also a link to `dbpedia:Para_State`[19] representing the State in which Xinguara is located. This way one can navigate from one resource to another, independently of which Linked Data repository each happens to be stored at.

Given the challenges of sharing and agreeing about meanings of predicates, the procedure of linking is not straightforward. Automatic linking can easily create inadequate links, but manual linking is often too time consuming [11]. A key research task is to support identity resolution, i.e., when two things denoted by two URI's are the same and when they are not. Linking also tends to have context-dependent outcomes. For example, information retrieval by a tourist can accept more loosely defined links (say, on partonomical relations) between places than retrieval for administrative tasks of authorities.

Specifying and publishing link types (i.e., predicates) encourages others to reuse them. For instance, the Citation Typing Ontology[20] [26] lists over 80 different types of citations (such as *cites as evidence*, *conforms* or *critiques*). If the GIScience community considers typing of citations between its publications, it will support deeper understanding of the impact of its work.

Furthermore, the sharing and reuse of spatial and temporal relations[21] as Linked Data by the community would support the large scale reuse of GIScience methods and applications. If two data sets use the same URI's for predicates and concepts, then queries and reasoning procedures tested with one data set will also work for the other.

10 New Challenges

With the adoption of Linked Data as a paradigm, new challenges emerge for research and practice. We briefly list some of them here, suggesting research directions for Geographic and other Information Sciences.

[18] See http://spatial.linkedscience.org/context/amazon/BRAZIL_MUNICIPALITY_1508407

[19] http://dbpedia.org/resource/Para_State

[20] http://purl.org/spar/cito

[21] Such as predicates defined by the Open Time and Space Core Vocabulary, see http://observedchange.com/tisc/ns/

1. How to deal with *raster data*: if one separates pure rasters from their interpretations into object concepts and treats both as Linked Data, what problems remain to be solved [24]?
2. How to deal with *time* in its many forms of relevance to linked data [19,28]?
3. How to exploit (recently standardized) RDF notions like multiple graphs for spatial data sets?
4. How to scope statement *validity* temporally [12] and spatially?
5. How to talk about statements themselves in a logically clean form, providing meta information (for a new and promising proposal, see [23]).
6. How to use such meta-statements in *trust and reputation* models [4]?
7. How to determine which *ontologies* are needed for geodata, how to reuse them, how to align them with other ontologies, and how to ensure community buy-in[22]?.
8. How to deal with *real-time* or near-real-time streams of data? [6]
9. How to extend Application Programming Interfaces (API's) to serve Linked Data, in addition to their typical CSV, JSON, and XML outputs [14]?
10. How do the large volumes of simply structured Linked Data affect *efficiency* in accessing and analyzing geographic data, in comparison with database systems and web services [16]?
11. How to better handle co-reference resolution to enable geo-data *conflation*?
12. Where are the hard limits, if any, of the triple data model for spatial data and which data should not be triplified (e.g., Well-Known-Text)?

The characteristics of the Linked Data paradigm, as described in this paper, provide a strong basis for addressing these and other challenges. In particular, the recognition that data are statements made by somebody somewhere at some time has already proven[23] to be one of the most powerful ideas when it comes to dealing with geographic information and when using spatial and temporal references as glues for information in general [15].

11 Conclusions

We discussed the paradigm shift afforded by Linked Data and Semantic Web technologies, highlighting its impacts on key questions of GIScience. Many of these impacts have to do with the changing role of database schemata, conventionally thought to be essential for modeling geographic data. This role needs to be revisited in the light of the triple model and the outsourcing of semantics into explicitly specified and shared vocabularies. Another set of impacts has to do with space and time as efficient integrators of data. The Linked Data approach makes this capacity explicit by enabling a global identification and publication of spatial and temporal references.

[22] These questions are being addressed by the so-called GeoVoCamps, see http://vocamp.org/wiki/Main_Page

[23] in projects such as http://lodum.de/life/

We argued that creating data in the form of Linked Data statements produces several benefits: statements can contradict and their validity can be time stamped, provenance information can be combined with the data itself, and semantics can be defined explicitly. With Linked Data, these benefits are built into the general Semantic Web infrastructure, creating a large-scale distributed (and spatially enabled) information infrastructure. We illustrated the Linked Data approach via examples ranging from mundane geographic facts through historical battles and environmental observations to bibliographic data. We ended with a list of a dozen research questions around novel challenges posed by Linked Data in GIScience.

The paper has focused on Linked Data, rather than Open or Linked Open Data. Yet, beyond the technical aspects discussed here, Linked Data has become an important vehicle for transparency in society. The paradigm of Open Government [21], popularized through national efforts in Brasil, the UK, the US, and other countries has rapidly spread and is reaching municipal levels in some countries [7]. With a wide range of available tools and a growing choice of vocabularies to convert data to Linked Data, anybody who wants (or is mandated) to open up geodata can and should now do so. Technical hurdles will no longer serve as an excuse to keep geodata hidden where there are no real reasons to do so. As for Open Data and Linked Data in general, the GIScience community has a great opportunity to help exploit location as an integrator across platforms, domains, and disciplines.

References

1. Dean Allemang and Jim Hendler. *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2008.
2. Robert Battle and Dave Kolas. Enabling the geospatial semantic web with parliament and geosparql. *Semantic Web*, 3(4):355–370, 2012.
3. Tim Berners-Lee. Linked Data – Design Issues. Online: http://www.w3.org/DesignIssues/LinkedData.html; last accessed 2014-02-03, 2009.
4. Mohamed Bishr and Werner Kuhn. Trust and Reputation Models for Quality Assessment of Human Sensor Observations. In *11th International Conference, COSIT 2013*, pages 53–73. Springer, Lecture Notes in Computer Science 8116, 2013.
5. Yaser Bishr. Overcoming the Semantic and Other Barriers to GIS Interoperability. *International Journal of Geographic Information Science*, 12(4):299–314, 1998.
6. Jean-Paul Calbimonte, H Jeung, Óscar Corcho, and K Aberer. Enabling query technologies for the semantic sensor web. *International Journal on Semantic Web and Information Systems*, 8(1):43–63, 2012.
7. Sergio Consoli, Andrea Giovanni Nuzzolese Aldo Gangemi, Valentina Presutti Silvio Peroni, Diego Reforgiato Recupero, and Daria Spampinato. Geolinked Open Data for the Municipality of Catania. In *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics*, Thessaloniki, Greece, June 2014. forthcoming.

8. Laura Diaz, Albert Remke, Tomi Kauppinen, Auriol Degbelo, Theodor Foerster, Christoph Stasch, Matthes Rieke, Bastian Schaeffer, Bastian Baranski, Arne Broering, and Andreas Wytzisk. Future SDI—Impulses from Geoinformatics Research and IT Trends. *International Journal of Spatial Data Infrastructures Research*, 7:378–410, 2012.
9. Max Egenhofer and Dominik Wilmsen. Changes in topological relations when splitting and merging regions. In *12th International Symposium on Spatial Data Handling*, pages 339–352, Vienna, Austria, 2006. Springer-Verlag.
10. Aldo Gangemi. Ontology Design Patterns for Semantic Web content. In *Proceedings of the Fourth International Semantic Web Conference*, pages 262–276. Springer, 2005.
11. John Goodwin, Catherine Dolbear, and Glen Hart. Geographical Linked Data: The Administrative Geography of Great Britain on the Semantic Web. *Transactions in GIS*, 12:19–30, 2008.
12. Claudio Gutierrez, Carlos Hurtado, and Ro Vaisman. Temporal RDF. In *Proceedings of the European Conference on the Semantic Web (ECSW2005)*, pages 93–107, 2005.
13. Harry Halpin, Patrick J Hayes, James P McCusker, Deborah L McGuinness, and Henry S Thompson. When owl:sameas isn't the same: An analysis of identity in linked data. In *The Semantic Web–ISWC 2010*, pages 305–320. Springer, 2010.
14. Glen Hart and Catherine Dolbear. *Linked Data: A Geographic Perspective.* Taylor & Francis, 2013.
15. Krzysztof Janowicz. The role of space and time for knowledge organization on the semantic web. *Semantic Web*, 1(1):25–32, 2010.
16. Jim Jones, Werner Kuhn, Carsten Keßler, and Simon Scheider. Making the web of data available via web feature services. In *Proceedings of the 17th AGILE Conference on Geographic Information Science*, Castellón, Spain, June 2014. forthcoming.
17. Tomi Kauppinen, Giovana Mira de Espindola, Jim Jones, Alber Sánchez, Benedikt Gräler, and Thomas Bartoschek. Linked Brazilian Amazon Rainforest Data. *Semantic Web Journal*, 5(2), 2014.
18. Tomi Kauppinen and Eero Hyvönen. *Modeling and Reasoning about Changes in Ontology Time Series*, pages 319–338. Integrated Series in Information Systems. Springer-Verlag, New York (NY), New York, NY, January 15 2007.
19. Tomi Kauppinen, Glauco Mantegari, Panu Paakkarinen, Heini Kuittinen, Eero Hyvönen, and Stefania Bandini. Determining relevance of imprecise temporal intervals for cultural heritage information retrieval. *Int. J. Hum.-Comput. Stud.*, 68(9):549–560, 2010.
20. Carsten Keßler, Krzysztof Janowicz, and Tomi Kauppinen. spatial@linkedscience Exploring the Research Field of GIScience with Linked Data. In Ningchuan Xiao, Mei-Po Kwan, Michael F. Goodchild, and Shashi Shekhar, editors, *Geographic Information Science*, volume 7478 of *Lecture Notes in Computer Science*, pages 102–115. Springer Berlin Heidelberg, 2012.
21. Daniel Lathrop and Laurel Ruma. *Open government: Collaboration, transparency, and participation in practice.* O'Reilly Media, Inc., 2010.
22. Timothy Lebo, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. Prov-o: The prov ontology. *W3C Recommendation, 30th April*, 2013.
23. Vinh Nguyen, Olivier Bodenreider, and Amit Sheth. Don't Like RDF Reification? Creating Meta Triples Describing Triples Using Singleton Property.

24. Thomas Scharrenbach, Sandro Bischof, Simon Fleischli, and Robert Weibel. Linked Raster Data. In Ningchuan Xiao, Mei-Po Kwan, Michael F. Goodchild, and Shashi Shekhar, editors, *Seventh International Conference on Geographic Information Science*, volume 7478 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2012.
25. Simon Scheider, Jim Jones, Alber Sanchez, and Carsten Keß ler. Encoding and querying historic map content. In *AGILE*, page in press, 2014.
26. David Shotton. CiTO, the Citation Typing Ontology. *Journal of Biomedical Semantics*, 1(Suppl. 1):S6, 2010.
27. Barry Smith and Berit Brogaard. Sixteen days. *The Journal of Medicine and Philosophy*, 28:45–78, 2003.
28. Johannes Trame, Carsten Keßler, and Werner Kuhn. Linked data and time–modeling researcher life lines by events. In *11th International Conference, COSIT 2013*, pages 205–223. Springer, Lecture Notes in Computer Science 8116, 2013.
29. Jun Zhao, Graham Klyne, and David Shotton. Provenance and linked data in biological data webs. In *The 17th International World Wide Web Conference (LDOW2008)*, 2008.