International Journal of Digital Earth

ISSN: 1753-8947 (Print) 1753-8955 (Online) Journal homepage: https://www.tandfonline.com/loi/tjde20

Geo-analytical question-answering with GIS

Simon Scheider, Enkhbold Nyamsuren, Han Kruiger & Haiqi Xu

To cite this article: Simon Scheider, Enkhbold Nyamsuren, Han Kruiger & Haiqi Xu (2020): Geo-analytical question-answering with GIS, International Journal of Digital Earth, DOI: 10.1080/17538947.2020.1738568

To link to this article: https://doi.org/10.1080/17538947.2020.1738568

© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

Published online: 12 Mar 2020.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tjde20

INTERNATIONAL JOURNAL OF DIGITAL EARTH
https://doi.org/10.1080/17538947.2020.1738568

OUTLOOK PAPER

OPEN ACCESS Check for updates

Geo-analytical question-answering with GIS

Simon Scheider, Enkhbold Nyamsuren, Han Kruiger and Haiqi Xu

Department of Human Geography and Spatial Planning, University Utrecht, Utrecht, The Netherlands

ABSTRACT

Question Answering (QA), the process of computing valid answers to questions formulated in natural language, has recently gained attention in both industry and academia. Translating this idea to the realm of geographic information systems (GIS) may open new opportunities for data scientists. In theory, analysts may simply ask spatial questions to exploit diverse geographic information resources, without a need to know how GIS tools and geodata sets interoperate. In this outlook article, we investigate the scientific challenges of geo-analytical question answering, introducing the problems of *unknown answers* and *indirect QA*. Furthermore, we argue why *core concepts of spatial information* play an important role in addressing this challenge, enabling us to describe analytic potentials, and to compose spatial questions and workflows for generating answers.

ARTICLE HISTORY
Received 26 July 2019
Accepted 1 March 2020

KEYWORDS
Geographic question answering; GIS; core concepts of spatial information; geo-analytics; geocomputation

1. Motivation

A large variety of analytical resources available on the Web and elsewhere offers genuine new opportunities for empirical scientists (Kitchin 2013). Take the example of a health scientist (Richardson et al. 2013). Millions of wearable sensors, exabytes of personal health records, billions of nodes on Open Street Map (OSM) and countless geolocated social media posts make it very probable that the spatial data needed for, say, finding out how environmental factors influence a person's health, stand ready for sophisticated ways of modeling.

Yet, the question of the health scientist can probably not *directly* be answered by this data. Data may not directly fit the purpose and may require further processing or analysis to generate a valid answer through geo-analytical tools. Geo-analytical tools, on the other hand, are difficult to employ when distributed across countless software programs. The 40 most well-known GIS software packages[1] together contain thousands of different tools (Ballatore, Scheider, and Lemmens 2018), and thousands of modules are added by online repositories such as PyPy[2] for Python or CRAN[3] for R. For the health scientist, it may simply take too much time to first learn GIS to find out whether it might answer his or her question using a particular data source.

What if the health scientist could simply ask a spatial question to find the right data and analysis functionality in an instant? This would tremendously reduce the effort needed, and it would also be a step towards exploiting geo-computational resources for analysts who are not GIS experts. Questions capture what analysts want to know and let them share and discuss their results independently from particular software or data formats (Vahedi, Kuhn, and Ballatore 2016). Yet, the kind of question-based analysis sketched here is unfortunately not possible today. Although Question-Answering

CONTACT S. Scheider sscheider@uu.nl Department of Human Geography and Spatial Planning, University Utrecht, 3584 CB Utrecht, The Netherlands

© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

(QA) systems have matured in recent times in both industry and academia (see Section 2.1), current approaches still focus on machine translations of questions into queries on factoid knowledge bases. To answer the question of our health scientist, however, it is not sufficient to query a database of known facts.

Let us illustrate this argument. We argue that the challenge of building a GIS that can directly answer questions (a question-based GIS) should *not* be conceived as a query task (as in ordinary QA), but rather as a particular *transformation task*. The question '*How far is it from Paris to Amsterdam?*' (Gao and Goodchild 2013) might be directly answered from a knowledge base, but it might also be translated into the concepts 'Spatial Distance',[4] 'Paris' and 'Amsterdam', which in turn can be easily mapped, for instance, to Google's routing function and corresponding nodes on Google Places. Consider, in contrast, the following question which is relevant in the context of Health Geography (Richardson et al. 2013):

> How much is Tom exposed to green space while running through Amsterdam?

First, it is unknown what the answer to this question is, as it depends on an analytic parameter (Tom's particular run). Furthermore, even if we grant that Tom's run is stored in a knowledge base, it is not obvious how an answer can be generated. For example, which combination of data sets and GIS tools as listed in Figure 1 would yield a valid answer? We call the latter kind of questions *analytic* in the following.

It is important to realize that GIS is fundamentally about *designing workflows* as answers to analytic questions for which answers are yet unknown. The work of GIS analysts, therefore, cannot be reduced to querying over known datasets. As an example, take the administrative Geography available on Wikidata[5]: Though Wikidata can easily answer the question in which state the US city San Diego is located, such kind of knowledge is of little interest for a Geographer.[6] Similar to statistics, GIS is a collection of methods for analysts, involving creativity in figuring out how data can be transformed to obtain an answer to a novel kind of question. What is needed is, therefore, a way to represent the *analytic potential* of a dataset for answering questions, based on a good theory about *spatial questions* as well as the possibilities of *operational transformations* provided by GIS.

We call this problem *geo-analytical QA*, which is part of a more general endeavour of *indirect QA* (Scheider, Ostermann, and Adams 2017). 'Indirect' means here that answers cannot be directly filtered out by queries, but need to involve transformations first. While indirect QA is relevant to all analytical sciences (including statistics and data science), not only to geographic information science, the inherent transformation possibilities and semantic constraints depend on the *kind of information* being transformed. This means that solutions will be specific to geographic information, and geo-analytical QA is special because transformations necessarily involve spatial concepts. In this

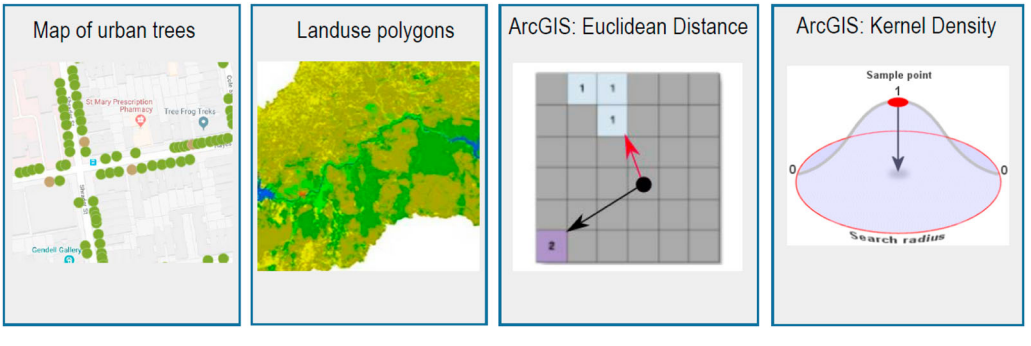


Figure 1. Which combination of tools and data would yield an answer to the question of the environmental health scientist? Example tools were taken from ArcGIS (http://desktop.arcgis.com/en/arcmap/). Data from the Amsterdam data portal (https://maps.amsterdam.nl).

outlook article, we make a case for geo-analytical QA as an autonomous research endeavour in the context of the Digital Earth, to which geospatial semantics (Janowicz et al. 2012), GIS workflow composition (Hofer et al. 2017) and spatial language processing (Hamzei et al. 2019) may contribute on an equal footing. We closely investigate this challenge and clarify the role that *core concepts of spatial information* could play in formulating and answering such types of questions. Since this is an outlook article, our discussion of solutions needs to remain preliminary.

2. The challenge: building a question-based GIS

In this section, we compare the task with current approaches in order to carve out its main challenges. This prepares the ground for arguing why semantic concepts in general are needed, and core concepts of spatial information in particular.

2.1. State of the art

Question answering has a long tradition in Artificial Intelligence (AI) and computational linguistics, going back to the wave of expert systems research of the twentieth century (Simmons 1970). Though these systems had limited impact, they touched upon many issues still relevant today, including linguistic grammars and semantic frames for parsing questions (Ofoghi, Yearwood, and Ma 2008). With the advent of the Web, a revival of QA systems occurred due to the availability of large query and answer sets (Lin 2002), with a potential to improve general information retrieval (IR) systems (Laurent, Séguéla, and Nègre 2006). Today, we distinguish knowledge-based Question Answering (KB QA), which derives answers from structured data (Diefenbach et al. 2018), and document-based QA, which finds answers within unstructured text (Kolomiyets and Moens 2011).

A recent review of the two categories of QA systems can be found in Shah et al. (2019). Compared to document-based QA systems that are more suitable for answering simple factoid questions, KB QA systems can answer questions that require reasoning over multiple factoids. While document-based QA systems can be rather easily generalized to different domains, KB QA systems require considerable manual effort in creating KBs and, thus, have limited cross-domain applicability. However, document-based QA systems require the availability of large corpora from which answers can be extracted. These disadvantages may indicate which QA system is more suitable in a particular case (Gupta and Gupta 2012). Recent efforts aim to create hybrid QA systems that combine elements of both (Mitra et al. 2019; Sawant et al. 2019). In case of GIS, such hybrid systems may be most suitable. Since GIS needs to answer complex analytical questions, KBs are really required, yet the domain also lacks a corpus of documents that can be leveraged.

The linked data cloud[7] and RDF[8] have been recently proposed for KB question-answering over data cubes (Höffner, Lehmann, and Usbeck 2016), allowing answers to be retrieved over many dimensions and resolution levels. The Semantic Web can be seen as a core technique for KB QA because its particular strength lies in reasoning over taxonomic concepts (Höffner et al. 2017), and large Web data bases such as DBpedia,[9] Yago[10] or WordNet[11] can be used for answer set generation (Bao et al. 2014). Main computational steps involve (1) the analysis of questions into phrases, (2) the mapping of phrases (including named entities) to the KB, (3) entity disambiguation, and the (4) construction and (5) firing of queries over the KB (Diefenbach et al. 2018).

Although introductory textbooks in GIS are focused around answering spatial questions (Heywood, Cornelius, and Carver 2011, ch.1), *geographic question answering* as such has only been subject of a limited number of research endeavors. In the past, researchers have proposed conversational and natural language interfaces to GIS (Cai et al. 2005). More recently, researchers have addressed how spatial questions can be translated to spatial query languages (Chen 2014; Pulla et al. 2013). Others also addressed how QA queries can be spatially and semantically *expanded* to arrive at a more successful answer rate (Mai et al. 2020). Gao and Goodchild (2013) matched *geo-analytical*

tools to questions based on keywords. More recently, Zhang et al. (2018) proposed a knowledge-based QA system for answering geographic questions that can be found in standardized tests of Chinese high school students. Overall, we need to ascertain a lack of effort in creating a system for addressing geo-analytical questions, which are so central to GIS.

Correspondingly, there is also a lack of studies on investigating these types of questions in particular. Few studies we have identified agree on informal categories (Heywood, Cornelius, and Carver 2011; Kraak and Ormeling 2013; O'Looney 2000; Allen 2016; Mitchell 2012), e.g. questions about relationships (e.g. What is the relationship between the local microclimate and locations of factories?) or questions about implications (e.g. If we build a new theme park here, what will be the effect on traffic flows?) (Heywood, Cornelius, and Carver 2011; Kraak and Ormeling 2013; O'Looney 2000; Mitchell 2012). While traditional QA systems also try to address questions about implication and relationship (Kolomiyets and Moens 2011; Wang 2006), it is assumed that answers can be directly retrieved from documents or logically inferred from knowledge bases. GIS commonly relies on analytic operations on raw data to answer these questions. Kuhn and Ballatore (2015) and Vahedi, Kuhn, and Ballatore (2016) revealed that the core concepts of spatial information, as suggested by Kuhn (2012), may play a central role in question formulation, as well as in describing these analytical operations.

2.2. Challenges in asking and answering geo-analytical questions

What is it that makes geo-analytical QA challenging? And why does the problem require more than what current QA technology has to offer? In a nutshell, since answers are basically unknown, they need to be given as workflows, which requires creativity in finding an answer. This, in essence, makes the problem a non-trivial learning task, since it requires going beyond matching of question and answer sets.

Creativity in answering. First, there is not only large degrees of freedom in how a spatial question can be asked, but also an equally large variety in how a given spatial question can be answered. The former problem is due to the flexibility of how a concept, such as distance, can be expressed in natural language ('How far?', 'How close?', 'what is the nearest …?'), leading to questions whose formulation is too far off the answer formulation in order to successfully match. Query expansion by increasing the tolerance of spatial query expressions (Mai et al. 2020) and also machine learning approaches have been used to circumvent this (Diefenbach et al. 2018; Bao et al. 2014). More importantly, however, there is usually also a certain *semantic ambiguity* in how a given expression in a question might be interpreted in terms of geospatial concepts. For example, the term 'green space' in the introductory question might equally well refer to an object (a park), a collection of objects (trees), or patches of a certain landcover category. This ambiguity is inherent to GIS and needs to be dealt with in geo-analytical QA. And a similar kind of ambiguity appears in *answer formulation*, too. There is no such thing as a most definite or probable answer contained in a geo-analytical KB. As a matter of fact, using a GIS, the same question can be answered in ways that are very different yet equally valid, based on combining different tools with different sorts of data sources (Chrisman 2002). For example, in order to address the question in Section 1, all of the tools in Figure 1 may render equally valid answers, using data sources ranging from Open Street Map (OSM) to official land-use statistics. The task is rather to capture the large variety of *valid* approaches to answering a given question, not of reducing an answer set to a most probable answer.

Question complexity. O'Looney (2000) suggests types of questions addressed by GIS and ranks them from simple to complex in the following order: location, condition, routing, pattern modeling, trend modeling, and what-if modeling. A 'location' question is a simple where question. An example of a condition question is 'What is the condition of the water treatment plant at 270 feet?'. An example of a pattern modeling question is 'What is the pattern of public spending in areas where the majority of residents are African American?'. While this particular typification of questions

may be contested, the acknowledgment of question complexity in GIS is of importance. Moreover, question types are not mutually exclusive. A condition question can also be a location question, and a pattern modeling question can also be a condition question. Therefore, answering a question in GIS requires its decomposition into simpler ones that need to be answered separately.

Indirect answers. The reason why geo-analytical questions are challenging in the first place is that their answers cannot be looked up. For those questions that QA systems usually handle, such as 'Who is the director of Forrest Gump?' (Bao et al. 2014), the answer is known. Our task is rather to match questions to answers which are unavailable, yet may be generated from what is known using analytic functions. This latter task requires capturing the *analytic potential of tools and data to answer a question*. In effect, this means to translate a spatial question into a query over a *transformation*: The query should match 'potential' datasets generated by some workflow (Figure 2), a novel computational challenge which was called 'indirect QA' in Scheider, Ostermann, and Adams (2017).

Non-trivial learning. All this makes geo-analytical QA a *non-trivial learning task*. On the one hand, the creativity and complexity involved, as well as the predominance of indirect answers, make it very hard to obtain representative training samples for questions and answers. On the other hand, training a QA system to give the most probable answer fails to capture precisely that a very different, maybe improbable, yet *valid* answer might be given.

2.3. The geo-analytical QA problem

One might argue that every indirect QA problem can be turned into an ordinary QA problem, simply by computing an answer and adding the answer to a database. However, such an approach would hardly solve the problem: it is simply impossible to precompute answers to every possible analytical question. It is for this reason that analytical QA really requires a new paradigm based on possible computational transformations, the latter being specific for GIS. Which steps would need to be taken for geo-analytical QA? As illustrated by Figure 2, there are three subproblems:

(1) Assessing the *analytic potential* of a geodata set. We cannot find out about the analytic potential of a dataset for a question by simply querying over it. Instead we need to assess whether a *transformation* of data exists which would answer a question. Similarly, we need to assess whether a certain GIS tool can be used in this transformation.
(2) Assessing the possibility of a transformation requires *synthesis of GIS workflows*. This generates possible answers. Workflows may be diverse but need to have a high quality, i.e. they need to be

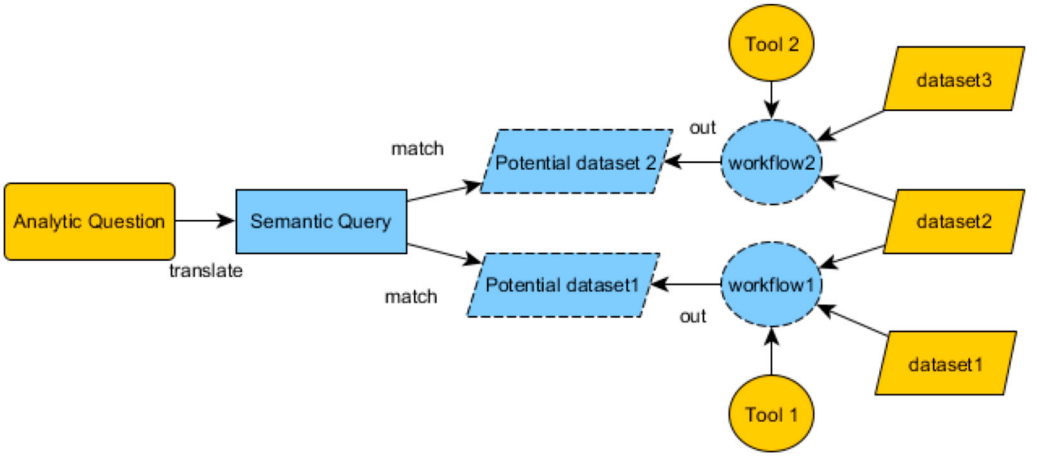


Figure 2. The problem of indirect QA. Blue boxes denote information which needs to be generated by an indirect QA system.

valid from a methodological point of view. This captures the creative aspect of geo-analytical QA in answering a question.

(3) Translating geo-analytical questions into *queries over such workflows*. To pick valid workflows as answers to a given question, we need to decompose the question into underlying concepts which match the outcome of some GIS-based transformation.

Restricting this to the case of geo-analytical questions is important: Our point is that all three problems are only solvable based on exploiting the semantic concepts that are contained in the questions and which are underlying GIS. This means that solutions need to be searched within the radius of geospatial semantics.

3. The role of core concepts

In the following, we suggest that core concepts of spatial information are indispensable not only for understanding how geo-analytical questions and answers are composed but also for knowing whether geodata is fit for answering. The reason is that they provide semantic constraints for posing spatial questions, as well as operational constraints for describing analytic potentials and for finding answers by constructing workflows.

3.1. Core concepts of spatial information

Core concepts of spatial information were proposed by Kuhn (2012), Kuhn and Ballatore (2015) as generic interfaces to GIS in the sense of conceptual 'lenses' through which the environment can be studied. Though they have been used in the sense of abstract data types (ADT),[12] they are considered results of human cognition and interpretation, and thus go beyond data types or formats. Core concepts in this latter sense were not invented by Kuhn, but are known and used implicitly by everybody who understands the essence of a GIS. Our task is to lift these concepts to an explicit level in order to make use of the information contained in them. Though a formal specification of core concepts is still ongoing work, and though the set of concepts have changed in the past to some extent,[13] there is a rather stable consensus of the following *content concepts*[14] on which we focus here:

- *Fields* are understood as particular kinds of functions (Galton 2004; Câmara, Freitas, and Casanova 1995; Scheider et al. 2016) whose domain are locations which allow for metric distance measurement, and whose range may be any kind of quality. Prime examples are temperature fields. A field function can change in time (Scheider et al. 2016). As quality values are separated by spatial distance, one can study change of a field as a function of spatial distance. Fields also offer the possibility of determining quality values at *arbitrary locations* in their domain. Missing quality values can therefore be estimated by *interpolation*. This concept is closely related to the notion of a field in physics (Einstein 1934).
- *Objects* are understood as entities which have a spatial region and diverse qualities that can change in time. This corresponds to the idea of 'endurants' in philosophy (Galton 2004), which can change their location and quality while remaining their identity. Objects are distinct from other concepts in the sense that they have identity (usually a name) and that they are fully localized in each moment of their existence, even if this location may be fuzzy. In this way objects give rise to trajectories, which are functions from time to location, and time series, which are functions from time to quality (Scheider et al. 2016). Geographic places, such as shopping malls or parking lots, are considered particular kinds of objects. Though they are not considered locations, they are localizable themselves.
- *Events* are understood as entities that, besides having identity and having qualities like objects, are not fully localized in each moment but *happen* during some time interval. Events thus correspond to particular kinds of 'occurrents' or 'perdurants' in Philosophy (Galton 2004). Since they have a

start and an end, they allow us to determine duration, and they might have objects, fields or spatial networks as participants. In GIS, we usually assume in addition that events are localizable similar to objects. Prime examples are earth quakes, having a time, a location as well as a magnitude.
- *Networks* are understood as quantified relations between objects. In this way, networks are able to measure a relationship between these objects. Similar to graphs, this relationship is quantified, e.g. in terms of an amount of flow or a distance. Networks in this sense are e.g. commuter flow matrices or distance edges in a road network.

Since core concepts have the mentioned properties, certain kinds of operations are naturally applied to them. For example, since fields are total functions on a metric space, their quality can be probed at every location and at every distance within that space, while object qualities cannot. Objects, on the other hand, can be counted, have spatial parts (mereology) and neighbors (topology), and furthermore give rise to sizes and closeness. Events can be ordered in time. Finally, since networks are relations between objects, they can always be projected to object qualities by fixing some source or destination, giving rise e.g. to catchments and service areas. Similar to levels of measurement and other semantic types (Chrisman 2002; Scheider and Huisjes 2019; Scheider and Tomko 2016), core concepts work as constraints to spatial analysis (Sinton 1978).

In contrast to ADTs, however, core concepts cannot be directly operated on. They rather work as spatial 'lenses' through which analysts see the geographic world when interpreting geodata (Vahedi, Kuhn, and Ballatore 2016). A given dataset, therefore, might be viewed with different lenses, making it possible to switch semantic perspectives and rendering interpretations inherently ambiguous. For example, the part of a road between two intersections can be regarded both as an object and a network element, and weather phenomena such as a storm can usually be viewed from a field perspective, an event perspective, as well as an object perspective. This ambiguity is part of GIS practice and thus needs to be taken into account. In Scheider et al. n.d.), we proposed *core concept data types* as a way to capture the different ways of how geodata types can represent core concepts. The ambiguity can be handled by allowing a dataset to be annotated by more than one core concept (see example in Section 3.2).

Note that our argument is not to convince people to use core concepts as a direct interface to a QA system. The idea is rather to use them as an internal representation of (possibly multiple) shared understandings of GIS question-answering resources. In the following, we argue for core concepts as part of (1) a type system for adding analytic potentials to data and tools, (2) an internal grammar for formulating questions and interpreting them as queries, and as (3) a way to compose answer workflows. These three aspects may provide the glue for solving geo-analytical QA.

3.2. The role of core concepts in describing analytic potentials of geodata sets

Though geodata types are commonly used in GIS, they are insufficient to assess the QA potential of geodata. The fact that raster and vector types do not capture the underlying concepts relevant for analysis is subject already of introductory GIS books. Figure 5(a), taken from Heywood, Cornelius, and Carver (2011), illustrates how diverse examples of spatial concepts (hotels, ski lifts, forst areas, roads and elevation surfaces) can be represented by both raster or vector data, and thus are orthogonal to these concepts. Furthermore, as Figure 3 demonstrates, the same geographic entities, such as points, can refer to different concepts such as objects, events, and field measurements. The arbitrariness of representing such concepts with geo data types indicates that core concepts add an independent but relevant piece of information to a given data type. Take the example of a 'forest' class in a landcover data set (Figure 5(b)). The fact that this data set is a polygon vector tessellation does not tell us much on how we can use it in analysis. In particular, it does not tell us that every polygon really represents a *homogeneous spot within a spatial field* of landcover values, and that therefore every location within a given polygon has the same landcover value. This way of representing a field was called *coverage* in Scheider et al. (2016), an example is the left map in Figure 4. A different

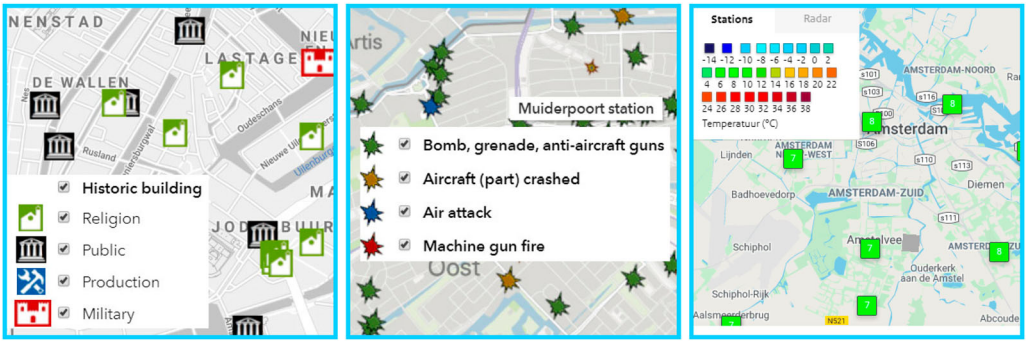


Figure 3. Point maps representing buildings as objects (left), war-time events (middle), and temperature field measurements (right). Data from the Amsterdam data portal (https://maps.amsterdam.nl).

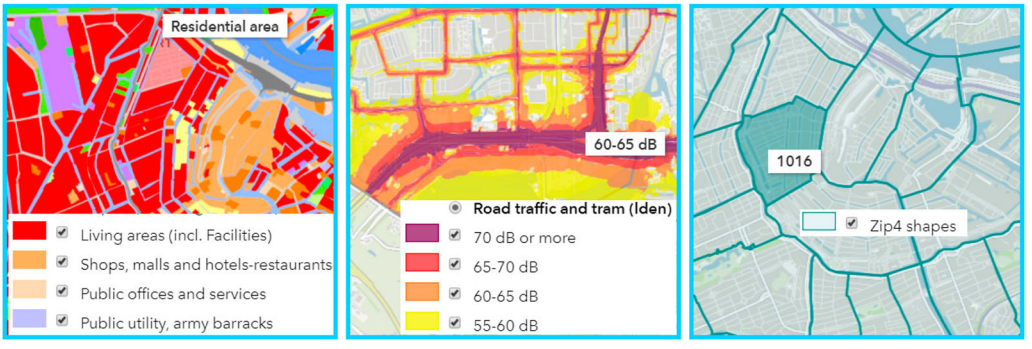
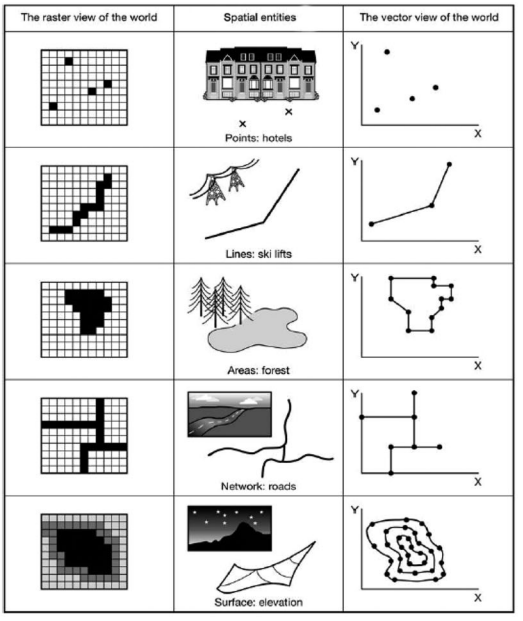


Figure 4. From left to right, polygon maps representing fields as coverages (land use types), contours (road traffic noise contour), and objects as lattices (zip code regions). Data from the Amsterdam data portal (https://maps.amsterdam.nl).

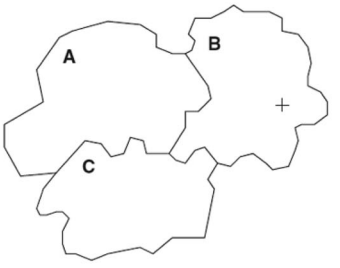
example of a vector based field representation is a *contour* map, as shown in the middle of Figure 4. However, a tessellated polygon data set may also be a representation of a tiled collection of spatial objects, such as in municipal statistics. For example, the right map in Figure 4 shows zip code regions. This way of representing objects was called *lattice* in Scheider et al. (2016).[15] Since municipalities are conceived as objects and not fields, their measured qualities, such as average elevation in a municipality, are valid only for the entire object, and not for any of its parts.

It is this distinction which largely influences how the dataset can be meaningfully analysed (Scheider and Tomko 2016; Scheider et al. 2016; Scheider, Ballatore, and Lemmens 2019). For example, *vector overlay* can be applied only to coverages, because it involves copying the measured quality for arbitrary parts within a given polygon. This is true also for raster overlay, where cell values are simply passed down to intersected parts of cells. Intersecting lattices, in contrast, requires areal interpolation instead (De Smith, Goodchild, and Longley 2007). In a similar way, point interpolation is applicable if and only if a point data set represents a field, and not a collection of objects or events (cf. Figure 3) (Scheider et al. 2016). This is opposed to density and distance computations, which require identifiable, countable and bounded spatial entities.

As these examples illustrate, there are plenty of reasons to assume that core concepts of spatial information, together with levels of measurement (Chrisman 2002) and related semantic distinctions (Scheider and Huisjes 2019), are indispensable in order to assess how a given data source might be transformed into meaningful answers. Our task in the future is therefore to (1) settle on a definite set of semantic types which capture the diverse ways how core concepts are represented within a given geodata type and (2) to find ways to scale up the semantic annotations across various geodata



(a)



	land cover type	average elevation (m)
A	Forest	631
B	Urban	220
C	Water	42
+	Urban	

(b)

Figure 5. Why core concepts add essential information to geodata types. (a) Concepts, such as fields, objects and networks, can always be represented by both geodata types, Vector or Raster. Source: Heywood, Cornelius, and Carver (2011). (b) The difference between representing a field (*coverage*) or an object (*lattice*) in terms of a vector tessellation. Land cover is an example for a coverage, and average elevation (or any other statistical aggregation) is an example for a lattice. The attribute located at the cross is determinable for the coverage, but not for the lattice (Scheider et al. 2016).

sources. Regarding the first problem, we have recently made a suggestion for a corresponding OWL[16]-based ontology pattern in Scheider et al. (n.d.). For the second task, machine learning (Scheider and Huisjes 2019) or crowd sourcing might be used (Khan et al. 2016).

3.3. The role of core concepts in posing geo-analytical questions

Core concepts also play an important role in formulating and interpreting questions. And similar to the data annotation task, there is often ambiguity in how to interpret a given question in terms of core concepts. In the following, we go through a range of example questions, highlighting different possible interpretations.

A question about spatial distance or spatial density usually implies spatial boundaries (to determine distances) and countability (to determine densities). Both are supplied by spatial objects. For example, such objects are parks and trees as in the following example:

How $^{(Field)}$far is the next $^{(Object)}$park in $^{(Object)}$Amsterdam?
How $^{(Field)}$densely are $^{(Object)}$trees located in $^{(Object)}$Amsterdam?

Distance to something and *density of something*, in turn, are interpreted here as spatial fields. While spatial distance itself is a quantified relation between locations, 'distance to something' projects this relation to a spatial field by fixing its range. Such a field is one way how we measure *exposure* in practice. Note that in all these cases, *Amsterdam* plays the role of another object whose region delimits the *extent*.

However, note that the same distance term might also be interpreted in terms of a *spatial network*. In this case, it is measured between pairs of objects, e.g. between any address and the next park object, and then projected to this address object:

How $^{(Network)}$$^{(Object)}$far is the next $^{(Object)}$park in $^{(Object)}$Amsterdam?

Questions may also directly query networks between two objects. The following question implies a network measuring runner flows on pairs of places (the latter understood as kinds of objects):

$^{(Network)}$How many runners run from $^{(Object)}$University campus $^{(Network)}$to $^{(Object)}$downtown?

Fields allow us to *probe* values at an arbitrary location within their extents ('How far away is the next park from *here*?'). If we interpret exposure to parks in terms of a distance field, then such locations may be supplied, e.g. by some spatial event (such as Tom's run), allowing us to ask:

$^{(Field)}$How much exposed to $^{(Object)}$parks is $^{(Event)}$Tom's run?

Tom's run may likewise be interpreted as a series of object locations used to do the probing:

$^{(Field)}$How much exposed to $^{(Object)}$parks is $^{(Object)}$Tom's run?

Furthermore, fields can be *summarized* into objects. In the following example, the term 'green' is interpreted as some homogeneous patch inside a field of land-use. This sub-field is summarized into the object of the Amsterdam municipality, where it constitutes a new object quality:

$^{(Object)}$How much is Amsterdam covered with $^{(Field)}$green?

Finally, in contrast to fields, objects can be spatially queried and counted using other objects:

How many $^{(Object)}$trees exist in $^{(Object)}$Amsterdam?

What these examples illustrate is not only that core concepts help decompose questions into (sub-) questions (e.g. the question about Tom's exposure can be decomposed into questions about a distance or density field), but also that parts of speech and goal concepts within questions can be interpreted as core concept transformations (e.g. the transformation of a land-use field ('green') into an object quality). Furthermore, even though such interpretation often allows for more than one decomposition, the resulting possibilities are constrained.

Every knowledge-based QA system needs a grammar that makes use of constraints to account for the space of possibilities to formulate questions (Diefenbach et al. 2018). We suggest that core concept transformations may provide the conceptual basis for a geo-analytical grammar, where a question's intent is composed of possible transformations. This grammar can be used in order to let users formulate questions which can then be automatically translated into meaningful queries over corresponding transformation workflows. As in the case of data annotations, the inherent ambiguity can be handled by allowing parts of speech to be parsed in terms of different concepts.

3.4. The role of core concepts in synthesizing answer workflows

In a certain sense, core concepts describe the origin of spatial information (Scheider et al. 2016). Correspondingly, they also provide necessary constraints for the applicability of functions to given information sources towards some geocomputational goal. For example, if we know that a data source represents a layer of objects, then this implies that it can be counted and thus spatial density can be computed. This idea can be exploited for loosely specifying and synthesizing GIS workflows (Lamprecht et al. 2010; Kasalica and Lamprecht 2018). To keep GIS workflow construction (Kind 2014) computationally manageable and to assure a sufficient workflow quality, automated program synthesis requires semantic constraints for a function's inputs and outputs which go beyond the current geodata types (Hofer et al. 2017). We suggest high-quality workflow synthesis may become possible using the semantic constraints that come with core concepts. For example, the two workflows in Figure 6 answer our question from Section 1. They were generated based on knowing that, similar to vector overlay, the ArcGIS tool 'Polygon to Raster' requires coverages as input (land-

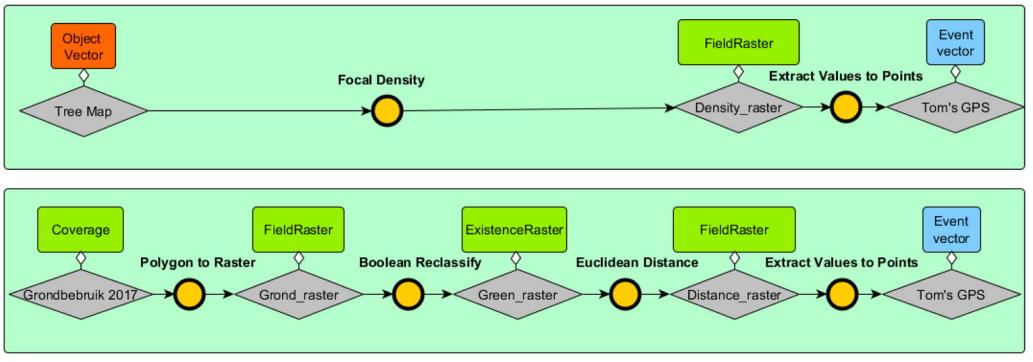


Figure 6. Exploiting core concepts for workflow composition to generate answers. Green implies field representations, red implies object representations, and blue implies event representations.

use data), while 'Focal Density' requires object representations (a map of trees). Note that both workflows generate field representations which are probed by Tom's GPS track.

The corresponding workflow specifications, consisting of input and output geodata types interpreted into core concepts, are given here. Note how the two answer workflow sequences satisfy these specifications,

Definition 3.1: How much is Tom exposed to green while running through Amsterdam?
in:ObjectVector, EventVector ∧ out:EventVector

resulting in the following workflow suggestion using core concept datatypes and ArcGIS operator names:

Workflow 1: ObjectVector, EventVector - Focal Density - FieldRaster - Extract Values to Points - EventVector

Note that the term 'green' was now interpreted into objects. Alternatively, one can interpret this term into a field representation, namely a certain land-use class (Coverage). This would look as follows:

Definition 3.2: How much is Tom exposed to green while running through Amsterdam?
in:Coverage, EventVector ∧ out:EventVector

Workflow 2: Coverage, EventVector - Polygon to Raster - FieldRaster - Boolean Reclassify - Existence Raster - Euclidean Distance - FieldRaster - Extract Values to Points - EventVector

Note while the difference in workflows reflects the ambiguity of semantic interpretation, some constraints are preserved. For example, the event vector used for probing the exposure field is, in both cases, origin as well as the goal of the workflow.

Future work should investigate the quality of such core concept based workflow generation algorithms. A first step has been made in Scheider et al. (n.d.), where workflows similar to the ones described here could be automatically generated based on *core concept data types*, specifying the goal and start concepts and using an extensive set of operational signatures of GIS functions in OWL.

4. Discussion and conclusion

We have argued that geo-analytical QA, that is, question-answering with a GIS, has a large potential for data science, yet seems a computational problem very different from ordinary

question-answering. While current approaches to geographic QA mainly rely on spatial queries on factoid knowledge bases, the main difference lies in the fact that geo-analytical knowledge bases usually do not contain answers but only references to analytical resources. Compared with a standard QA setting, geo-analytical QA therefore requires accounting for the creativity of analysis, and for assessing the potential of data sources and tools to answer a given question (indirect QA). Its relevance lies in spatial questions occurring in all data sciences, while data scientists are less and less able to learn the variety of functions and data that would allow them to answer their questions. It should also be noted that indirect QA is not a problem unique to the geo-spatial domain. For example, statistics face similar problems of overwhelming variety of tools and data. However, the scope of our study is limited to geo-analytical problems. Given this scope, we have further argued and illustrated with examples why core concepts are essential to handle this task. First, they provide many of the needed semantic constraints that capture the analytic potential of tools and data sources beyond current data types. Second, they are essential in interpreting and posing spatial questions, in the sense that core concepts are used to construct and fill the semantic roles in a query. And third, the constraints implied by core concepts can be exploited for workflow construction in order to compute possible answers, along the lines of Kasalica and Lamprecht (2018). In order to realize such a geo-analytical QA system, future work should focus on describing the analytic potential of tools and data. Semantic typing is needed to capture core concepts across different data representations (Scheider, Ballatore, and Lemmens 2019). Empirical research is necessary to identify the roles of core concepts in spatial question patterns. A corresponding grammar would allow us to translate a spatial question into a query. To investigate the feasibility of answer computation, workflow synthesis methods need to be tested on annotated resources, and their answering potential needs to be measured by matching queries to workflows using a transformation language based on core concepts. The theoretical and technical implications of this research may be of relevance not only to the practical application of GIS but also to GIS education and training. Any solution to the geo-analytical QA problem may be turned into a handbook or a manual for GIS analysts. Such manual would assist in developing necessary expert skills for decomposing geographic questions in terms of core concepts and for mapping them to geodata sets and tools to formulate answer workflows.

Notes

1. Including ArcGIS, PostGIS, QGIS, ERDAS, Grass GIS, R, etc.
2. https://pypy.org/
3. https://cran.r-project.org/
4. Though 'distance' is usually defined operationally, we consider it nevertheless a concept. 'Distance' can be represented both in terms of data and operations and can be even rendered vague in language.
5. https://www.wikidata.org
6. While these kinds of spatial questions are dominating current search engines such as Bing, cf. Hamzei et al. (2019), they are usually not of much interest for geographic studies.
7. https://lod-cloud.net/
8. https://www.w3.org/RDF/
9. http://dbpedia.org/
10. https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/
11. https://wordnet.princeton.edu/
12. In computer science, an abstract data type (ADT) is an abstraction of a type of data in terms of its *behavior*, i.e. in terms of the applicability of operations to the data (Liskov and Zilles 1974).
13. http://spatial.ucsb.edu/core-concepts-of-spatial-information/
14. Kuhn (2012) distinguishes content concepts from quality concepts (resolution and accuracy) and base concepts, such as location.
15. Note that the term *lattice* as used here does *not* refer to the corresponding mathematical concept. It rather goes back to the notion in spatial statistics.
16. https://www.w3.org/TR/owl2-overview/

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 803498 (QuAnGIS)).

References

Allen, David W2016. *GIS tutorial 2: spatial analysis workbook*. Esri Press.
Ballatore, Andrea, Simon Scheider, and Rob Lemmens.. 2018. "Patterns of Consumption and Connectedness in GIS Web Sources." In *The Annual International Conference on Geographic Information Science*, 129–148. Springer.
Bao, Junwei, Nan Duan, Ming Zhou, and Tiejun Zhao.. 2014. "Knowledge-Based Question Answering as Machine Translation." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, 967–976.
Cai, Guoray, Hongmei Wang, Alan M. MacEachren, and Sven Fuhrmann. 2005. "Natural Conversational Interfaces to Geospatial Databases." *Transactions in GIS* 9 (2): 199–221.
Câmara, Gilberto, U Freitas, and Marco Antônio Casanova. 1995. "Fields and Objects Algebras for GIS Operations." In *Proceedings of III Brazilian Symposium on GIS*, 407–424.
Chen, Wei. 2014. "Developing a Framework for Geographic Question Answering Systems Using GIS, Natural Language Processing, Machine Learning, and Ontologies." PhD diss., Ohio State University.
Chrisman, Nicholas. 2002. *Exploring Geographic Information Systems*. 2nd ed. New York: Wiley.
De Smith, Michael John, Michael F Goodchild, and Paul Longley. 2007. *Geospatial Analysis: A Comprehensive Guide to Principles, Techniques and Software Tools*. Leicester, UK: Troubador Publishing Ltd.
Diefenbach, Dennis, Vanessa Lopez, Kamal Singh, and Pierre Maret. 2018. "Core Techniques of Question Answering Systems over Knowledge Bases: A Survey." *Knowledge and Information Systems* 55 (3): 529–569.
Einstein, Albert. 1934. "On the Method of Theoretical Physics." *Philosophy of Science* 1 (2): 163–169.
Galton, Antony. 2004. "Fields and Objects in Space, Time, and Space-Time." *Spatial Cognition and Computation* 4 (1): 39–68.
Gao, Song, and Michael F. Goodchild. 2013. "Asking Spatial Questions to Identify GIS Functionality." In *2013 Fourth International Conference on Computing for Geospatial Research and Application*, 106–110. IEEE.
Gupta, Poonam, and Vishal Gupta. 2012. "A Survey of Text Question Answering Techniques." *International Journal of Computer Applications* 53 (4): 0975–8887.
Hamzei, Ehsan, Haonan Li Maria Vasardani, Timothy Baldwin, Stephan Winter, and Martin Tomko.. 2019. "Place Questions and Human-Generated Answers: A Data Analysis Approach." In *The Annual International Conference on Geographic Information Science*, 3–19. Springer.
Heywood, Ian, Sarah Cornelius, and Steve Carver. 2011. *An Introduction to Geographical Information Systems*. 4th ed. Harlow, UK: Pearson Prentice Hall.
Hofer, Barbara, Stephan Mäs, Johannes Brauner, and Lars Bernard. 2017. "Towards a Knowledge Base to Support Geoprocessing Workflow Development." *International Journal of Geographical Information Science* 31 (4): 694–716.
Höffner, Konrad, Jens Lehmann, and Ricardo Usbeck. 2016. "CubeQA–Question Answering on RDF Data Cubes." In *International Semantic Web Conference*, 325–340. Springer.
Höffner, Konrad, Sebastian Walter, Edgard Marx, Ricardo Usbeck, Jens Lehmann, and Axel-Cyrille Ngonga Ngomo. 2017. "Survey on Challenges of Question Answering in the Semantic Web." *Semantic Web* 8 (6): 895–920.
Janowicz, Krzysztof, Simon Scheider, Todd Pehle, and Glen Hart. 2012. "Geospatial Semantics and Linked Spatiotemporal Data–Past, Present, and Future." *Semantic Web* 3 (4): 321–332.
Kasalica, Vedran, and Anna-Lena Lamprecht. 2018. "Automated Composition of Scientific Workflows: A Case Study on Geographic Data Manipulation." In *2018 IEEE 14th International Conference on e-Science (e-Science)*, 362–363. IEEE.
Khan, Vassilis Javed, Gurjot Dhillon, Maarten Piso, and Kimberly Schelle.. 2016. "Crowdsourcing User and Design Research." In *Collaboration in creative design*, 121–148. Springer.
Kind, Josephine. 2014. "Creation of Topographic Maps." In *Process Design for Natural Scientists*, 229–238. Springer.
Kitchin, Rob. 2013. "Big Data and Human Geography: Opportunities, Challenges and Risks." *Dialogues in Human Geography* 3 (3): 262–267.
Kolomiyets, Oleksandr, and Marie-Francine Moens. 2011. "A Survey on Question Answering Technology from an Information Retrieval Perspective." *Information Sciences* 181 (24): 5412–5434.

Kraak, Menno-Jan, and Ferdinand Jan Ormeling. 2013. *Cartography: Visualization of Spatial Data*. Harlow, UK: Routledge.
Kuhn, Werner. 2012. "Core Concepts of Spatial Information for Transdisciplinary Research." *International Journal of Geographical Information Science* 26 (12): 2267–2276.
Kuhn, Werner, and Andrea Ballatore. 2015. "Designing a Language for Spatial Computing." In *AGILE 2015*, 309–326. Springer.
Lamprecht, Anna-Lena, Stefan Naujokat, Tiziana Margaria, and Bernhard Steffen.. 2010. "Synthesis-Based Loose Programming." In *2010 Seventh International Conference on the Quality of Information and Communications Technology*, 262–267. IEEE.
Laurent, Dominique, Patrick Séguéla, and Sophie Nègre. 2006. "QA Better than IR?" In *Proceedings of the Workshop on Multilingual Question Answering*, 1–8. Association for Computational Linguistics.
Lin, Jimmy J. 2002. "The Web as a Resource for Question Answering: Perspectives and Challenges." In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, Canary Islands, Spain, 1–8.
Liskov, Barbara, and Stephen Zilles.. 1974. "Programming with Abstract Data Types." In *ACM Sigplan Notices*, Vol. 9 (4), 50–59. ACM.
Mai, Gengchen, Bo Yan, Krzysztof Janowicz, and Rui Zhu.. 2020. "Relaxing Unanswerable Geographic Questions Using a Spatially Explicit Knowledge Graph Embedding Model." In *Geospatial Technologies for Local and Regional Development*. Springer.
Mitchell, Andy. 2012. *Modeling Suitability, Movement, and Interaction*. Redlands, CA: Esri Press.
Mitra, Arindam, Peter Clark, Oyvind Tafjord, and Chitta Baral.. 2019. "Declarative Question Answering over Knowledge Bases containing Natural Language Text with Answer Set Programming." arXiv preprint arXiv:1905.00198.
Ofoghi, Bahadorreza, John Yearwood, and Liping Ma. 2008. "The Impact of Semantic Class Identification and Semantic Role Labeling on Natural Language Answer Extraction." In *European Conference on Information Retrieval*, 430–437. Springer.
O'Looney, John. 2000. *Beyond Maps: GIS and Decision Making in Local Government*. Redlands, CA: ESRI, Inc.
Pulla, Venkata S. K., Chandra S Jammi, Prashant Tiwari, Minas Gjoka, and Athina Markopoulou.. 2013. "QuestCrowd: A Location-Based Question Answering System with Participation Incentives." In *2013 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 75–76. IEEE.
Richardson, Douglas B., Nora D. Volkow, Mei-Po Kwan, Robert M. Kaplan, Michael F. Goodchild, and Robert T. Croyle. 2013. "Spatial Turn in Health Research." *Science* 339 (6126): 1390–1392.
Sawant, Uma, Saurabh Garg, Soumen Chakrabarti, and Ganesh Ramakrishnan. 2019. "Neural Architecture for Question Answering Using a Knowledge Graph and Web Corpus." *Information Retrieval Journal* 22 (3-4): 324–349.
Scheider, Simon, Andrea Ballatore, and Rob Lemmens. 2019. "Finding and Sharing GIS Methods Based on the Questions They Answer." *International Journal of Digital Earth* 12 (5): 594–613.
Scheider, Simon, Benedikt Gräler, Edzer Pebesma, and Christoph Stasch. 2016. "Modeling Spatiotemporal Information Generation." *International Journal of Geographical Information Science* 30 (10): 1980–2008.
Scheider, Simon, and Mark D. Huisjes. 2019. "Distinguishing Extensive and Intensive Properties for Meaningful Geocomputation and Mapping." *International Journal of Geographical Information Science* 33 (1): 28–54.
Scheider, Simon, Rogier Meerlo, Vedran Kasalica, and Anna-Lena Lamprecht.. n.d. "Ontology of Core Concept Data Types for Answering Geo-Analytical Questions." http://josis.org/index.php/josis/article/viewArticle/555.
Scheider, Simon, Frank O. Ostermann, and Benjamin Adams. 2017. "Why Good Data Analysts Need to Be Critical Synthesists. Determining the Role of Semantics in Data Analysis." *Future Generation Computer Systems* 72: 11–22.
Scheider, Simon, and Martin Tomko. 2016. "Knowing Whether Spatio-Temporal Analysis Procedures are Applicable to Datasets." In *FOIS*, 67–80.
Shah, Asad Ali, Sri Devi Ravana, Suraya Hamid, and Maizatul Akmar Ismail. 2019. "Accuracy Evaluation of Methods and Techniques in Web-Based Question Answering Systems: A Survey." *Knowledge and Information Systems* 58 (3): 611–650.
Simmons, Robert F. 1970. "Natural Language Question-Answering Systems: 1969." *Communications of the ACM* 13 (1): 15–30.
Sinton, David. 1978. "The Inherent Structure of Information as a Constraint to Analysis: Mapped Thematic Data as a Case Study." *Harvard Papers on Geographic Information Systems*.
Vahedi, Behzad, Werner Kuhn, and Andrea Ballatore. 2016. "Question-Based Spatial Computing–A Case Study." In *Geospatial Data in a Changing World*, 37–50. Springer.
Wang, Mengqiu. 2006. "A Survey of Answer Extraction Techniques in Factoid Question Answering." *Computational Linguistics* 1 (1): 1–14.
Zhang, Zhiwei, Lingling Zhang, Hao Zhang, Weizhuo He, Zequn Sun, Gong Cheng, Qizhi Liu, Xinyu Dai, and Yuzhong Qu. 2018. "Towards Answering Geography Questions in Gaokao: A Hybrid Approach." In *China Conference on Knowledge Graph and Semantic Computing*, 1–13. Springer.