Patterns of consumption and connectedness in GIS Web sources

Andrea Ballatore, Simon Scheider, and Rob Lemmens

AGILE Conference 2018 - Lund, Sweden, 12-15 June
Author copy

Abstract Every day, practitioners, researchers, and students consult the Web to meet their information needs about GIS concepts and tools. How do we improve GIS in terms of conceptual organisation, findability, interoperability and relevance for user needs? So far, efforts have been mainly top-down, overlooking the actual usage of software and tools. In this article, we critically explore the potential of Web science to gain knowledge about tool usage and public interest in GIScience concepts. First, we analyse behavioural data from Google Trends, showing clear patterns in searches for GIS software. Second, we analyse the visits to GIScience-related websites, highlighting the continued dominance of ESRI, but also the rapid emergence of Web-based new tools and services. We then study the views of Wikipedia articles to enable the quantification of methods and tools' popularity. Fourth, we deploy web crawling and network analysis on the ArcGIS documentation to observe the relevance and conceptual associations among tools. Finally, in order to facilitate the study of GIS usage across the Web, we propose a linked-data inventory to identify Web resources related to GI concepts, methods, and tools. This inventory will also enable researchers, practitioners, and students to find what methods are available across software packages, and where to get information about them.

Key words: GIS; Geographic information science; GIScience; GIS operations; Web science; Google Trends; Wikipedia; ArcGIS; Linked data

Andrea Ballatore (✉)
Department of Geography, Birkbeck, University of London, London, UK
e-mail: a.ballatore@bbk.ac.uk

Simon Scheider
Human Geography and Planning, Universiteit Utrecht, Utrecht, NL
e-mail: s.scheider@uu.nl

Rob Lemmens
Fac. of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, NL
e-mail: r.l.g.lemmens@utwente.nl

1 Introduction

The Web offers invaluable resources for researchers, practitioners, and students of geographic information science (GIScience) and spatial data science. To meet their information needs, users search and consume online information originating from technical manuals, software documentation, academic websites, blogs, forums, discussion boards, as well as social media. The same set of GIScience ideas, ranging from core concepts (Kuhn, 2012) to methods such as buffer and interpolation, are found in a vast number of heterogeneous, incompatible software suits, such as ArcGIS, QGIS, R, and Carto. Hundreds of GIS tools exist, and it is currently not known how much and when they are actually used. Analysing usage patterns would be immensely useful to improve the conceptual organisation, usability, and findability of these tools, as well as the methods and concepts that underpin them. Knowledge about to what extent GIS software, tools, and methods attract the attention of users would be valuable to ground research in this direction: Researchers, developers, and practitioners could relate their work to information needs in a data-driven way.

GIS users would benefit from a mapping between tools, concepts, and Web pages that describe them. For example, spatial analysts could grasp the workings of methods at an abstract level across software, identifying suitable tools more effectively. Teachers could indicate to students the variety of ways in which similar concepts and methods are implemented in real software packages. Software developers could better integrate their products to existing software, making their tools more findable and better linked to the GIScience concepts that they use. Several initiatives aimed at structuring GIS concepts, methods, and tools with a rather top-down approach, only observing the tools and their formal definitions, without considering behavioural data (Lemmens, 2006; Gao and Goodchild, 2013; Kuhn and Ballatore, 2015; Scheider et al., 2017).

In this article, we take the Web as a resource to study the patterns of consumption of GIS-related information, focussing on tools, software packages, organisations, as well as more abstract GIScience concepts. By adopting a Web science approach (Hendler et al., 2008), this study focusses on the following research questions:

- To what extent are Web sources useful to study GIS usage?
- What is the relative popularity of GI tools and organisations?
- How are tools associated with each other?
- What is the popularity in GIS methods and concepts?
- How can we connect Web resources to GIScience concepts and methods using linked data?

After reviewing existing efforts in understanding and mapping GIScience usage (Section 2), we report on this study in five parts, organized as follows. First, Google Trends data about GIS is explored critically in Section 3. A pool of highly visible GIS-related websites is studied in Section 4, relying on data from Web analytics firms Alexa Internet and SimilarWeb. Section 5 then focusses on the popularity of Wikipedia articles related to GIScience, charting topics that attract high, medium,

and low interest. Subsequently, Section 6 performs a network analysis of the most visited GIS website, i.e., the documentation of ArcGIS. As a way to improve the organisation and findability of these resources, we then outline the proof-of-concept of a linked-data inventory, which highlight commonalities and relationships across these GIScience resources (Section 7). As part of this study, we also tested NLP methods, such as topic models (Blei, 2012), on a corpus of GIScience websites, but as these did not seem to yield interesting results, we left them out of this article. Finally, Section 8 draws conclusions and directions for future work.

As part of our efforts to make GIS more semantically structured, all the resources created as part of this study are available in an online repository as open knowledge.[1]

2 Related work

GIScience principles are in use in a plethora of tools. Currently, we face a lack of up-to-date knowledge on which GI tools exist, how they link to each other and to underlying core concepts (Kuhn and Ballatore, 2015). This is necessary to know how tools should best be used in a given context (Hofer et al., 2017), and how we can translate between GIS workflows (Bernard et al., 2014; Ludäscher et al., 2006), abstracting from particular software packages (Hinsen, 2014; Scheider and Ballatore, 2018). Currently, all we have is a vague idea about different GIS software products and their associated (and often closed) worlds of terminology (Steiniger and Hunter, 2013). Better linkage would have positive effects in both GIScience practice and education.

The World Wide Web constitutes a network of resources that can be exploited for Web science (Hendler et al., 2008) and, more generally, for data-driven science (Hey et al., 2009). Its wealth of inter-connected, distributed, user-generated content makes it an obvious candidate for studying usage patterns of informational resources and tools (Castellano et al., 2013), on a scale which is unprecedented and may be impossible to reach with traditional usability or empirical user studies (Kveladze et al., 2013).

Empirical studies in GIScience that make use of the Web and social media to explore human behaviour abound. They include estimating the location of tweeting users (Hecht et al., 2011), or harvesting geospatial information about places from social media feeds (Stefanidis et al., 2013; McKenzie et al., 2015) and from text corpora (Hollenstein and Purves, 2010), and are based on mature, well-established methods (Ferrara et al., 2014). New approaches for extracting semantic information from unstructured texts (Blei, 2012; Ramage et al., 2009) have been used to describe and link information resources about GI tools and methods (Hu et al., 2015; Gao and Goodchild, 2013). Web statistics derived from search engines like Google can inform researchers across disciplinary boundaries (Stephens-Davidowitz, 2013).

[1] `https://github.com/simonscheider/GISTrends`

Yet, it is debatable to what extent the "unstructured" Web can be a reliable empirical resource for estimating GIScience content consumption patterns. Foundational critique about the big data hype was raised in recent years (Boyd and Crawford, 2011), addressing the representational bias in human language texts on the Web (Caliskan et al., 2017), which can lead to severe estimation errors in a data-driven science (Lazer et al., 2014). Furthermore, the missing structure of Web information, i.e., the lack of "semantic rails for the data train", were recently criticised (Janowicz et al., 2014), making it hard to pre-select data and tools in a way that accounts for their inherent biases, and thus to separate signal from noise (Scheider et al., 2017). The linked data paradigm may offer a strategy to counter this weakness of pure bottom-up methods, in so far as it provides an infrastructure for sharing unstructured as well as structured and semantically precise information about tools and data (Brauner, 2015; Hofer et al., 2017; Scheider and Ballatore, 2018), including the classification of GIS functions. Besides the informal classifications in several GIS text books, a few efforts have presented approaches for formally classifying GIS functions (Albrecht, 1998; Lemmens, 2006; Brauner, 2015).

A strategy for integrating bottom-up and top-down approaches to research on GI usage is still lacking (Scheider et al., 2017), and a critical exploration of GIScience online resources is overdue. Hence, in this study, we deploy a Web science approach to inspect what online information about GIScience and GI tools is consumed. What follows is a first mapping of GIScience online, based on behavioural data from a number of complementary sources, assessing their usefulness and reliability.

3 GIS software tools on Google Trends

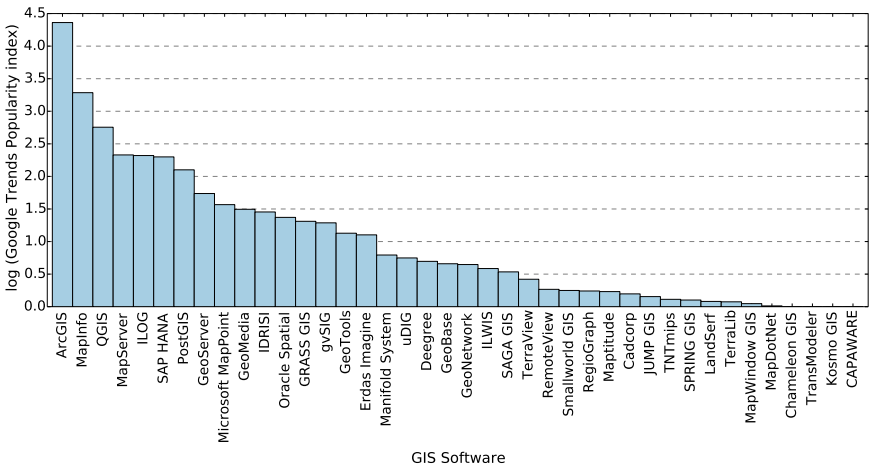
Google Trends[2] offers aggregate search statistics generated in the Google ecosystem, and is a valuable source for studying the behaviour of users on the Web, for example to predict economic patterns (Choi and Varian, 2012), analyse consumer behaviour (Goel et al., 2010), and explore cultural changes (Stephens-Davidowitz, 2013). The service provides relative search frequencies for arbitrary terms in a weekly resolution since 2004. Results are aggregated per country, and are given as an index where 100 denotes the highest frequency measured for the given terms over time. A maximum of five terms can be compared against one another.[3] Since the volume is given only as a relative index from 100, and Google rounds off volumes that are below a certain resolution threshold, term frequencies can easily drop to zero. For this reason, the selection of comparable terms is essential for this method to provide interpretable results.

Since GIS users commonly rely on the Web as an information resource to find out about software, tools, methods and their intended usage based on the Google search engine, relative volume of searches for GIScience-related keywords and topics

[2] `https://trends.google.com`

[3] `https://medium.com/@pewresearch/using-google-trends-data-for-research-here-are-6-questions-to-ask-a7097f5fb526`

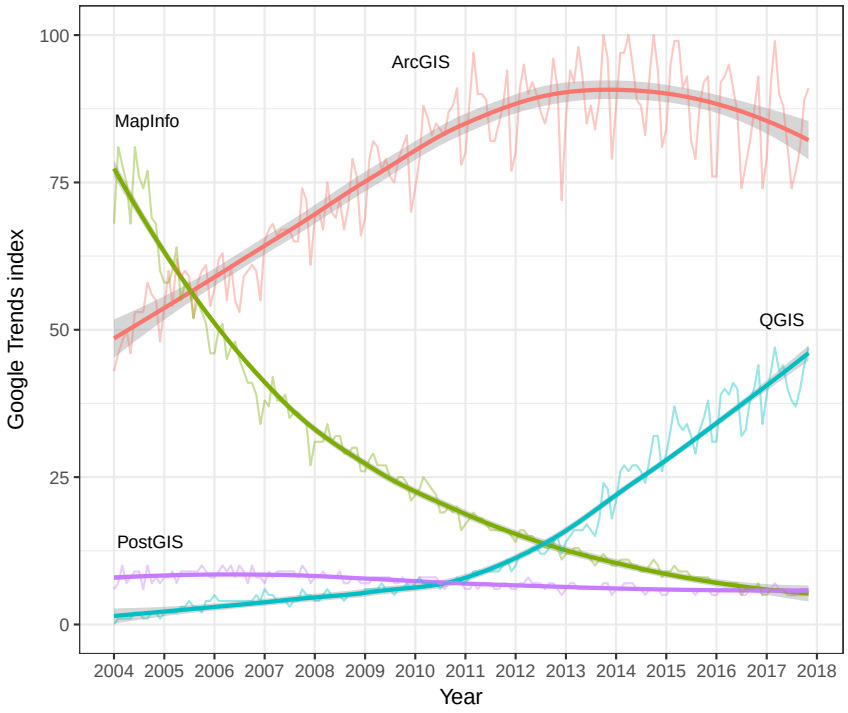
Fig. 1 Relative popularity of GIS software names on Google Trends, compared with the search term "ArcGIS" and averaged over the entire Google Trend history (2004-2017). On the y axis, we show the natural logarithm of the popularity index compared among five terms including "ArcGIS" during the entire period. Data collected in November 2017.



provide an indicator for the prominence of topics and tools. In this section, we focus on searches for GIS tools, selecting their official names as keywords for software packages and tools which we gathered from the Web as described in Section 7. To collect Google Trends data, we devised a method that selects four keywords at a time against a reference keyword with comparably high volume, averaging relative trends over the entire recording period (from 2004 to 2017). This way, it becomes possible to compare a larger set of keywords, circumventing the problem that search volumes are not provided as absolute numbers. To ensure the interpretability of the results, we only search for individual keywords, and not for topics, i.e., aggregates of keywords identified by Google.

Figure 1 displays an averaged relative search volume index over all GIS software tools, measured against the reference term "ArcGIS", since this term was used most often. We used a logarithmic scale because search volume differs a lot between terms. Note that we had to exclude the term "AutoCAD", because its search volume is several magnitudes higher than that of any GIS tool, making the comparison difficult. Furthermore, in the case of polysemic names that coincide with frequent search terms like "Grass", we added the string "GIS" to restrict the search to the desired semantic field. Results appear meaningful, suggesting that ArcGIS is most the popular GIS tool, followed by MapInfo, and QGIS. PostGIS, Intergraph's GeoMedia, and GeoServer have a considerably lower but still comparable search volume, while tools like the deegree (sic) map server and ILWIS obtain much lower online attention. Similarly, exactly 4 tools are in fact too infrequently searched to be comparable with the reference term.

Fig. 2 Relative popularity of selected GIS software product names over the entire Google Trends history (2004-2017). The trend lines are produced with a LOESS regression. Data collected in November 2017.



As illustrated in Figure 2, the temporal trends for these tool names clearly show that, while QGIS started to grow rapidly in 2011, searches for MapInfo have been continuously decreasing since 2004. More surprisingly, interest in ArcGIS enjoyed robust growth until 2014, and then levelled off and started to decline.

While trends for software products yield meaningful results, this is unfortunately not the case for the GIS tools. We carried out the same kind of comparison on all ArcGIS tools contained in the popular toolboxes "Spatial Analyst", "Conversion Tools", and "Analysis Tools", compared with the reference term "ArcGIS". On the surface, it seems that some tool names are very frequently searched. At closer inspection, however, these tool names are highly polysemic. The most searched toolnames are "Aggregate", "Corridor", "Watershed" and "Visibility" with an index greater than or equal to 50. However, it is apparent that these terms have meanings beyond GIScience, and therefore the results bear large amounts of noise. Searches for "Table To Excel" might be popular for reasons entirely unrelated to GIS, and therefore cannot say anything about the usage of the ArcGIS tool of this same name. Adding software names to these tool names (e.g., "ArcGIS Aggregate") delimits the semantic field correctly, but the low volume of searches makes all scores fall to zero.

A similar problem arises when searching for more general GIScience topics, such as the term “Kriging”.

In summary, Google Trends analysis works fairly well with unambiguous, distinctive, and relatively popular terms (e.g., “ArcGIS”, “QGIS”, “Kriging”), but is utterly unusable for more polysemic terms used in many semantic contexts, such as “join”, “buffer”, and “interpolation”. Apart from mainstream tools, other searches appear to be too infrequent to identify discernible signals.

4 GIScience top websites

As the Web is a prominent locus of information production and consumption, in this section we investigate which websites offer GIScience-related information and quantitatively observe their popularity. In this analysis, the data is collected from two sources: Alexa Internet is a US-based online marketing company that collects detailed statistics on online resources.[4] SimilarWeb is a London-based company that offers analogous web analytics resources.[5] These companies gather a variety of indicators of online behaviour to estimate the traffic to websites along different facets, including spatial, temporal, and demographic variables. Taking website *wikipedia.org* as an example, Alexa Internet states that it is the fifth most visited website worldwide.[6] Along the same lines, SimilarWeb estimates that it is the eleventh most visited website, with about 6.6B visits per month.[7] In most instances, Alexa Internet produces rankings that are significantly higher than those by SimilarWeb. This data can be used to quantify the engagement of audiences with websites, and observe trends in web-based consumer behaviour.

To draw a picture of GIScience content online, we selected a pool of websites based on the tools discussed in Section 3, starting from a Wikipedia-based list of GIS tools. To broaden the scope beyond tools, we included a range of specialist magazines (GIS Geography and GIM International), and a set of notable organisations that produce online content related to GIScience (e.g., the Open Geospatial Consortium and the Ordnance Survey). All these websites contain GIScience-related content, including product descriptions, software documentation, tutorials, examples, and discussions. While this pool cannot be exhaustive in its current form, we believe it captures a significant portion of top online content that most GIS practitioners and students consult.

From a methodological perspective, the data provided by Alexa Internet and SimilarWeb present limitations. The websites operate as black boxes, and it is hard to ascertain the accuracy of the estimates. Moreover, the data does not provide statistics about subsets of websites, limiting the analysis to websites that are thematically

[4] `https://www.alexa.com`

[5] `https://www.similarweb.com`

[6] `https://www.alexa.com/siteinfo/wikipedia.org`

[7] `https://www.similarweb.com/website/wikipedia.org`

focussed. For example, it is possible to obtain data about *stackexchange.com*, but not about *gis.stackexchange.com*, which would be more relevant to this study. Similarly, several software tools do not have dedicated domains, but are hosted at large repositories. For example, the software package PySAL is hosted on *readthedocs.io* (*pysal.readthedocs.io*), and it is hard to obtain traffic statistics. For this reason, many potentially relevant sub-domains had to be excluded. That said, we consider this data to be sufficient as an indication of the magnitude of online popularity of these resources.

We collected engagement information for a pool of 55 GIScience-related websites, of which 18 were discarded for lack of data. Table 1 summarises the results of this analysis: For each website, the table indicates the average worldwide rank calculated from the ranks from the two sources, thus reducing bias. In the interest of brevity, the ranks and visit counts were heavily rounded to the thousands or millions. To the best of our knowledge, other websites that we initially considered ranked more than six millionth on either platform, without enough data to produce estimates, and were therefore removed. The table also includes the SimilarWeb estimate of monthly visits, and not unique visitors, i.e. the same web user can generate more than one visit.

The top countries indicated by Alexa Internet are selected based on the absolute volume of visits, hence countries with large populations tend to dominate. The US, China, and India are top countries for most websites, with some notable exceptions, e.g., Italy, Algeria, and other countries for specific websites. A set of important, but non-thematically specific, websites about technologies like Oracle, Python, and R is included at the end of table, also providing a reference point for the GIScience websites. The pool of websites is available on the GitHub repository, and can be re-used for similar analyses.

Unsurprisingly, the websites of ArcGIS and ESRI emerge as the most popular in the pool, with about 19M monthly visits, and ranking between 5,000th and 19,000th in the world. Another traditional GIS, MapInfo by Pitney Bowes, also maintains a popular position, but it is hard to estimate visits specific to the tool, and not to other branches of the company. More interestingly, emergent competitors to ESRI are visible, including aggressive Web start-ups Mapbox and Carto, which attract respectively 2.9M and 724,000 monthly visits. Free and open source GI tools (Steiniger and Hunter, 2013) reach high visibility, spearheaded by desktop-based QGIS (1.4M monthly visits). Web mapping JavaScript libraries Leaflet and OpenLayers have become extremely popular since the late 2000s. Mature tools, such as GDAL, GeoTools, PostGIS, MapServer, and GeoServer, obtain between 230,000 and 50,000 monthly visits, suggesting persistent engagement by their communities of users. The other websites obtained lower ranks and visits, and are therefore not discussed in detail.

Table 1 Popular GIScience websites, according to Alexa Internet and SimilarWeb data, as of 15 November 2017. The ranks are a measure of global popularity of the websites. The visits are a monthly estimate from SimilarWeb for October 2017. (*) These entries include non-spatial content, and therefore tend to rank higher. For example, *oracle.com* covers all Oracle products, not just Oracle Spatial and Graph.

Product / organisation	Website	Average rank	*Alexa rank*	*SimWeb rank*	Monthly visits	Alexa top countries
ArcGIS	arcgis.com	5K	3K	6K	14.2M	US, Canada
ESRI	esri.com	13K	8K	19K	4.9M	US, China
Mapbox	mapbox.com	16K	16K	16K	2.9M	US, China
Ordnance Survey	ordnancesurvey.co.uk	44K	55K	34K	1.3M	UK, US
QGIS	qgis.org	45K	32K	58K	1.4M	US, China
Leaflet	leafletjs.com	57K	50K	65K	1.5M	US, China
Carto	carto.com	60K	42K	77K	724K	US, Spain
OS GEO, GRASS GIS	osgeo.org	99K	68K	131K	640K	US, China
OpenLayers	openlayers.org	110K	76K	141K	371K	China
GIS Geography	gisgeography.com	162K	85K	240K	372K	US, India
Intergraph	intergraph.com	203K	136K	270K	164K	US, India
PostGIS	postgis.net	214K	147K	281K	231K	US, Belgium
GDAL	gdal.org	237K	135K	340K	180K	China, Japan
GeoServer	geoserver.org	270K	172K	367K	130K	Brazil, US
Erdas Imagine	hexagongeospatial.com	286K	203K	370K	175K	US, South Africa
OGC	opengeospatial.org	392K	268K	516K	132K	US, Spain
MapServer	mapserver.org	516K	353K	679K	79K	Italy, India
GIM International	gim-international.com	596K	329K	864K	67K	India, US
Directions Magazine	directionsmag.com	608K	268K	947K	57K	US, India
GeoTools	geotools.org	658K	499K	817K	57K	Algeria, Russia
Geography UK	geography.org.uk	806K	546K	1.1M	52K	UK
uDIG	udig.refractions.net	960K	960K	–	–	US, India
TurfJS	turfjs.org	992K	616K	1.3M	–	US, India
gvSIG	gvsig.com	1M	635K	1.4M	–	Spain, Russia
Spatial.ly	spatial.ly	1.1M	1M	1.3M	–	–
GeoNetwork	geonetwork-opensource.org	1.2M	943K	1.4M	–	Germany
TerrSet, formerly IDRISI	clarklabs.org	1.5M	520K	2.5M	–	India
52 North	52north.org	1.7M	750K	2.6M	–	–
Cadcorp	cadcorp.com	2.2M	1.2M	3.2M	–	India
Manifold System	manifold.net	2.3M	1.3M	3.3M	–	US
R Spatial	r-spatial.org	3.5M	2.3M	4.7M	–	–
OpenJump	openjump.org	3.6M	2M	5.2M	–	–
Deegree	deegree.org	4.1M	2.8M	5.3M	–	–
Oracle*	oracle.com	685	369	1K	63.4M	US, India
Python*	python.org	2K	879	3K	35.1M	US, China
R*	r-project.org	11K	7K	15K	7.2M	US, China
Pitney Bowes, MapInfo*	pitneybowes.com	43K	28K	59K	1.9M	US, UK

5 GIScience content in Wikipedia

In our mapping of GIScience Web resources, we dedicate particular attention to Wikipedia, which represents without doubt a prominent entry point to much Web content. Wikipedia articles are highly heterogeneous, and cover from very general

(e.g., geography) to very specific technical topics, such as Moran's *I*. For this analysis, we selected a pool of Wikipedia articles in English that are related to GIScience. As GIScience is by its nature a multi-disciplinary, porous domain, we selected a very broad range of topics by crawling the website from a set of highly central seed pages,[8] and collecting the links to other articles for two edges in the network. This procedure generated a list of 1,073 articles, which we manually scanned and classified as either GIScience-related or not. For example, we included *location intelligence* and *contour line*, while *personal computer* and *animal cognition* were discarded as only marginally relevant to this analysis. When in doubt, we included the article, recognising the degree of subjectivity in this classification.

As a result of this process, we obtained a list of 349 relevant pages. In this analysis we focus exclusively on page views, and not on other indicators, such as number of edits and article length. Because of its constrained structure, Wikipedia articles are thematically delimited, and page views provide an indicator of interest in a given topic. However, the data has indeed known limitations that should not be ignored. The page view counts are sensitive to current events that can generate short-lived bursts of views, as well as to polysemy, when pages with unrelated topics with some of the same keywords are opened by mistake. Links on the main page of Wikipedia can also boost views without other explanatory factors.[9] In sum, we consider these problems acceptable in our set of GIScience-related articles, which – alas – do not seem to obtain mainstream visibility on the Web. For each page, we retrieved usage statistics from the Wikimedia API, focussing on monthly views from October 2016 to October 2017.[10] The average monthly views were then calculated as a proxy of interest in the article topics.

In the set of the 349 pages, the number of monthly views ranges from 15 to about 117,000, with a median of 1,055. To provide context to this data, the most popular 50 pages in Wikipedia currently obtain between 6.9M and 1M monthly views.[11] As expected in hypertext-based data, the distribution is heavily skewed towards a small set of pages that attract most of the views, with a tail of low-traffic pages. The top 10% of the pages generate about 65% of the total views in the set. Table 2 shows a summary of this analysis, ordering the Wikipedia articles by monthly views. Based on Jenks natural breaks, we grouped the pages into five classes, ranging from very high volume of views to very low. Some articles in the last group were omitted for the sake of brevity. The complete table can be found on the GitHub repository.

The most visited articles, having more than 42,000 views per month, include geographic coordinate systems, GPS, GIS, latitude, and longitude. The difference in interest between GIS and GIScience is staggering, with respectively 71,000 and 2,200 views, suggesting that, while GI *systems* keep attracting a broad audience, GI *science* remains a small academic discipline, particularly when compared with its

[8] Seed pages include *Geographic information science*, *Category:Geographic information systems*, *List of geographic information systems software*, and *Geoinformatics*.

[9] https://en.wikipedia.org/wiki/Wikipedia:Pageview_statistics#Accuracy_of_the_tools

[10] https://wikitech.wikimedia.org/wiki/Analytics/AQS/Pageviews

[11] https://tools.wmflabs.org/topviews/

Table 2 Wikipedia articles about GIScience-related topics, grouped by number of monthly views. In each group, the articles are sorted in descending order by monthly visits. Different colours are used to denote a concept, a software, a tool and an organisation. Please note that some articles are referred to with more than one title, obtaining different views (e.g., *Global Positioning System* and *GPS*). The prefix for the pages is *https://en.wikipedia.org/wiki/*.

Monthly views (thousands)	*Wikipedia articles [size of group]*
Very high [42,120)	Geographic coordinate system, Global Positioning System, R (programming language), Geographic information system, Geography, Latitude, Map, Cluster analysis, Data science, Longitude, Topology [11]
High [20,42)	Surveying, Census, Map projection, Remote sensing, Crowdsourcing, Cartography, SAP HANA, Tessellation, Universal Transverse Mercator coordinate system, Ontology (information science), Raster graphics, Human geography, Data visualization, Contour line, OpenStreetMap, Data model [16]
Medium [6,20)	Satellite navigation, Garmin, Geotechnical engineering, National Geospatial-Intelligence Agency, Aerial photography, Geomorphology, Geotagging, Satellite imagery, Geodesy, ArcGIS, Heat map, Scale (map), Spatial analysis, GPS, Geoid, Geophysics, Kriging, Digital elevation model, TomTom, Geodetic datum, R-tree, Quadtree, Political geography, List of geographic information systems software, Choropleth map, Geolocation, Location-based service, Esri, Ordnance Survey, Public Land Survey System, History of geography, Well-known text, Thematic map, Bing Maps, QGIS, Cadastre, Geohash, Citizen science, Gazetteer, Wikimapia, Geomatics, Spatial database, GeoJSON, Web Mercator, Cultural geography, Landsat program, Geospatial analysis [47]
Low [2,6)	Outline of geography, Geospatial intelligence, Geoinformatics, GIS file formats, Spatial reference system, Lambert conformal conic projection, Geographical distance, Google Sky, Inverse distance weighting, Moran's I, ISO 19115, Maps, Maidenhead Locator System, LIDAR, Geography Markup Language, ISO 10005, Ingres (database), Development geography, Geostatistics, Google Moon, Georeferencing, List of GIS data sources, What3words, Geolocation software, Scientific visualization, GIS, SVG, GeoTIFF, Regional geography, Population geography, Jenks natural breaks optimization, MapInfo Professional, Virtual globe, Crime mapping, Image rectification, Triangulated irregular network, WGS84, Web Feature Service, USGS, List of programs for point cloud processing, PostGIS, Datum (geodesy), Big Data, Philosophy of geography, CartoDB, Erdas Imagine, ArcMap, GDAL, GRASS GIS, Meridian arc, Geographic information science, Global Map, Geodynamics, Cartographer, Behavioral geography, Orthogonal projection, GeoServer [57]
Very low [.01,2)	Health geography, DE-9IM, Global navigation satellite system, Geodemographic segmentation, WikiMapia, Minimum bounding rectangle, Geographic profiling, Geovisualization, Modifiable areal unit problem, Urban informatics, ArcGIS Server, Spatial index, Web Coverage Service, Data model (GIS), GPS receiver, Cartographic generalization, British national grid reference system, Geomarketing, Spatiotemporal database, Simple Features, Location intelligence, Grid (spatial index), Environmental geography, Vector Map, Polygons, Treemap, Satellite geodesy, MrSID, Land administration, ArcInfo, Georeference, Geoportal, SpatiaLite, Volunteered geographic information, Spatial query, USGS DEM, Data Mining, Geocode, Vector tiles, CityEngine, Counter-mapping, NAD83, Indicators of spatial association, Buffer (GIS), Mapnik, Oracle Spatial and Graph, GeoMedia, Geographic information systems, MapInfo Corporation, GIS and public health, Viewshed, Digital Earth, GvSIG, GeoSPARQL, SAGA GIS, Cartographic relief depiction, ... [218]

cognate disciplines of geography (69,000 views) and data science (52,000 views). Similarly, crowdsourcing remains a highly consulted article (30,000 views), while the more specific volunteered geographic information (VGI) is a niche topic, with only 1,000 monthly views.

Unlike the GIScience websites covered in Section 4, the articles in this analysis show how Wikipedia tend to have good coverage of topics at a high level of abstraction (e.g., thematic map) and software packages (e.g., QGIS), but minimal inclusion of GI methods, such as a buffer and weighted overlay. This helps explaining why the ESRI and ArcGIS websites still take the lion's share of GIScience online traffic. We hope that the data reported in this analysis can help GIScience practitioners and students guide efforts to make the discipline more visible online, increasing the coverage, connectedness, and quality of GIScience-related articles.

6 The structure of the ArcGIS documentation

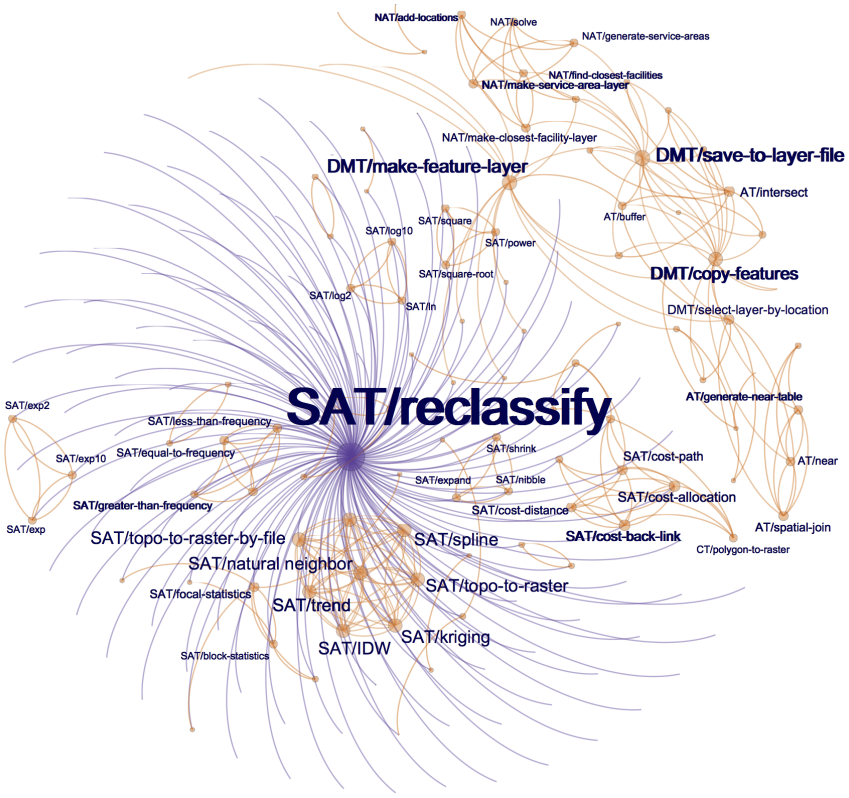
The online documentation of ArcGIS is the most visited GIS-related website (see Section 4), and therefore offers the opportunity of studying a software tool in detail. First, we scraped the website *arcgis.com*, collecting 928 documentation pages about the popular software suite. These pages include tool documentation, tutorials, and various forms of technical explanations, mixing applied and scientific content. As visible in the ArcGIS graphical interface, the tools are grouped in arbitrary toolboxes, such as the Spatial Analyst. The documentation describes different versions of the tools, and therefore, to avoid duplication, we restricted the analysis to a popular major version (10.x), for a total of 285 pages about tools. For example, the popular buffer tool is documented in a Web page.[12]

To observe the semantic associations between the tools, we run a network analysis on the tool-related pages, aiming at identifying which tools tend to be used together. A manual inspection of the links shows a rather sparse network, without clearly interpretable, non-trivial patterns. Hence, we perform a *graph selection* (Stell and Worboys, 1999) which connects pages through at most one intermediate page. That is, paths between pages $p_0 \rightarrow p_1 \rightarrow p_2$ correspond to a second-order edge $p_0 \rightarrow p_2$ in the resulting graph. The second-order edges between ArcGIS tool Web pages are summarised in Figure 3. Meaningful patterns start emerging when the second-order graph is further cleaned from obvious hubs, such as toolbox and tutorial pages that are highly inter-linked. A link from one tool to another tool means here that there is either a direct Web link between corresponding tool Web pages, or over one intermediate page, where the latter can also be a non-tool page (e.g., a page describing general principles of the software). Node and label sizes are scaled relative to the node degree in the network.

It is possible to see in this network that there are several tools acting as central nodes. The node with the highest degree is *Reclassify* from the Spatial Analyst

[12] `http://desktop.arcgis.com/en/arcmap/10.3/tools/analysis-toolbox/buffer.htm`

Fig. 3 Second-order links between ArcGIS tool Web pages, showing their degree in terms of node size and colour from orange (low) to blue (high), taken from the toolboxes Spatial Analyst (SAT), Data Management (DMT), Network Analyst (NAT), Analysis (AT), Conversion (CT), Geocoding (GT). The network layout was obtained with the Fruchterman-Reingold algorithm.



toolbox (SAT), with an in-degree of 142, followed by *Save-to-layer-file* (18), *Make-feature-layer* (17), *Copy-features* (15) from the Data Management toolbox (DMT) (see Table 3). The node centrality and connectivity pattern reveals an insight: In raster analysis, the *Reclassify* tool is actually a central means to transform a raster layer based on its cell values. It therefore acts as an interface between all kinds of raster tools, such as map algebra operations. This tool has other tools pointing to it, but does not point itself to other pages (see out-degree in Table 3).

Furthermore, layer operations from the Data Management toolbox are central for all kinds of GIS workflows to deal with layers as inputs and outputs. Lastly, one can see a meaningful cluster containing the spatial analyst tools *Kriging*, *Trend*, *Spline*, *IDW* and *Topo-to-raster* and *Natural Neighbor*. These are tools that can be used to interpolate surfaces in Digital Elevation Models (DEM). Even smaller subclusters are nicely interpretable, such as the cluster of *Cost-Distance*, *Cost-Back-Link*, *Cost-*

Table 3 Node degrees in the second-order graph of ArcGIS tool Web pages.

Tool	Degree	Out-degree
SAT/reclassify	142	0
DMT/save-to-layer-file	18	2
DMT/make-feature-layer	17	2
DMT/copy-features	15	0
SAT/idw	15	8
SAT/spline	15	8
SAT/topo-to-raster	15	8
SAT/spline-with-barriers	15	8
SAT/trend	15	8
SAT/topo-to-raster-by-file	15	8
SAT/kriging	15	8
SAT/natural-neighbor	15	8
SAT/cost-allocation	9	6
SAT/cost-back-link	9	6

Allocation, which together form a set of highly interdependent tools for least cost path analysis on cost surface raster layers. Note also that clusters partially overlap with and link different toolboxes. This method can be used to analyse connections between tools, making implicit knowledge emerge from the website network.

7 Linked inventory of GIS tools

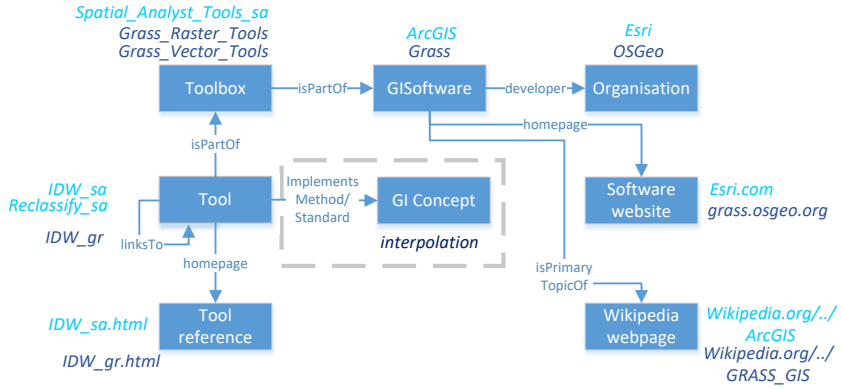
To systematize studies of these Web resources and to share our results about usage patterns of GIS software, tools and concepts, we suggest a way to unambiguously describe and identify the involved resources with linked data. For this purpose, we designed a comprehensive linked inventory that describes GIS tools and their implementations across different packages (e.g., ArcGIS, GRASS, and R).[13] This dataset was used as a basis for all Web analyses performed in previous chapters, and contains resources derived as a result of this study. To generate the inventory, an initial set of GIS software packages was identified from Wikipedia articles,[14] and then enriched with links from DBPedia.[15] For example, in Listing 1, ArcGIS is described with standard RDF vocabularies.

[13] `http://geographicknowledge.de/vocab/GISTools.ttl[.rdf]`

[14] `https://en.wikipedia.org/wiki/Comparison_of_geographic_information_systems_software`

[15] See for example `http://dbpedia.org/page/ArcGIS`

Fig. 4 Overview of the linked data inventory. The boxes represent the tool vocabulary, while examples of GIS tools are in italic. The dashed area represents future work.




```
dbp:ArcGIS a dbo:Software;
  dbo:developer dbp:Esri;
  foaf:homepage <http://www.esri.com/software/arcgis>;
  foaf:isPrimaryTopicOf <https://en.wikipedia.org/wiki/ArcGIS>;
  foaf:name "ArcGIS".
```

Listing 1 Describing GIS software products using linked data.

To obtain information about the tools contained in each software, we additionally scraped manuals on the Web, for example that of GRASS GIS.[16] When possible we used scripts within a given package to generate tool inventories, linking them to a preliminary subset of software packages based on our Web study, e.g., *arcpy* in ArcGIS. For example, Listing 2 shows how we used `dct:isPartOf` from Dublin Core terms to nest tools within toolboxes and packages such as ArcGIS. Finally, we enriched this dataset with tool network information scraped from web texts and their hyperlinks (see Section 6). This enabled us to link tools to webpages (using `foaf:homepage`) and to encode their network structure (with the SIOC term `sioc:links_to`) into linked data. Figure 4 shows a schematic representation of the linked data inventory as also described in Listing 1 and 2. The linked data approach facilitates the interconnection of tools and their descriptions and can form the basis for further connections with GI concept definitions in text books, tutorials, curricula, etc.

[16] `https://grass.osgeo.org/grass72/manuals/keywords.html`

```
@prefix tools:<http://geographicknowledge.de/vocab/GISTools.rdf#>.
@prefix sioc:<http://rdfs.org/sioc/ns#>.
@prefix dct:<http://purl.org/dc/terms/>.
@prefix foaf:<http://xmlns.com/foaf/0.1/>.

tools:Spatial_Analyst_Tools_sa a gis:Toolbox;
  dct:isPartOf dbp:ArcGIS;
  rdfs:label "Spatial Analyst Tools(sa)".
tools:IDW_sa a gis:Tool;
  dct:isPartOf tools:Spatial_Analyst_Tools_sa;
  foaf:homepage <http://desktop.arcgis.com/.../idw.htm>;
  sioc:links_to tools:Reclassify_sa.
```

Listing 2 Capturing GIS tools, toolboxes, websites and Web links as linked data.

Once extended beyond this proof-of-concept, we hope that this resource will support education and research purposes, becoming a basis for further research on GIS tools usage patterns.

8 Conclusions

In this article, we explored the Web science approach to gather new knowledge about the consumption of online information about GIS tools, software, and concepts. As part of our efforts to improve the conceptual organisation of GIS, we critically examined Google Trends data about the popularity of tools, and the top websites that host GIScience content, based on publicly available Web-analytics data. Subsequently, we studied two notable websites, including behavioural data about Wikipedia articles and the network structure of the ArcGIS online documentation. Based on this study, we designed the structure of a linked-data inventory, which connects these Web resources across GIS software, tools, and concepts, and presented examples of its use.

In sum, the Web scientific approach allowed us to discover patterns buried in behavioural and structural aspects of websites, producing some interesting findings. Google Trends allows granular tracking of software popularity, confirming the dominance of ESRI products, but also the emergence of new tools and companies. Alexa Internet and SimilarWeb enable the estimation of visits to GIS websites. In Wikipedia, we can observe the popularity over time of a plethora of topics, ranging from software to scientific concepts and methods. Our analysis also suggests much higher popularity for term “GIS” as opposed to “GIScience”, potentially directing efforts to better represent the discipline online. Finally, the network analysis of online documentation allowed us to capture meaningful functional relationships between tools that are not immediately apparent, and which may be used as a basis to recommend tools.

However, this study also highlighted several limitations of Web science. Noise caused by semantic ambiguity of keywords limits the interpretability of some analyses, particularly in the case of Google Trends. Moreover, this approach focused on large-scale online information consumption, which is at best a proxy to user

behaviours, such as GIS usage and adoption. The latter can only be measured in a direct way based on traditional research methods, such as local log files, surveys and interviews, which are restricted to a small scale. In this sense, access to corporate data would be immensely beneficial to understand tool usage (but unlikely to happen). Finally, we realized that the Web science approach is heavily dependent on what software organisations and the majority of users deem relevant, and this may just not what an analyst needs in a particular situation.

For future research, we envisage several worthwhile directions. It is paramount to produce more structured information about the relevance of GIS tools, methods, and concepts, boosting the precision and recall of user searches (Ballatore et al., 2016), instead of relying on unstructured data such as texts. For this reason, the inventory we outlined in this article should be incrementally extended to reach broader coverage of existing tools, embedding them into a coherent conceptual framework. Furthermore, to support data scientists and students, we must increase the semantic depth of our inventory, capturing the functionality of tools and related concepts (Scheider and Ballatore, 2018), which is only partially possible with the Web scientific method. This would result in a better linkage between methods (e.g., buffer and interpolation) and their software implementations, for example in R and ArcGIS.

Finally, in order to map GIS software, tools, and related websites, more comprehensive analyses are needed, increasing the completeness of our mapping with input from the GIScience community. For this purpose, crowdsourcing would facilitate information gathering and error-correction, supporting the iterative revision of our assumptions. A near-complete, maintainable set of tools, software, and websites will allow researchers and practitioners to find suitable resources, monitoring the evolution of this broad technical landscape.

References

Albrecht, J. (1998). Universal analytical GIS operations: A task-oriented systematization of data structure-independent GIS functionality. In Onsrud, H. and Craglia, M., editors, *Geographic information research: Transatlantic perspectives*, pages 577–591. Taylor and Francis, New York.

Ballatore, A., Kuhn, W., Hegarty, M., and Parsons, E. (2016). Spatial approaches to information search. *Spatial Cognition & Computation*, 16(4):245–254.

Bernard, L., Mäs, S., Müller, M., Henzen, C., and Brauner, J. (2014). Scientific geodata infrastructures: challenges, approaches and directions. *International Journal of Digital Earth*, 7(7):613–633.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Boyd, D. and Crawford, K. (2011). Six provocations for Big Data. In *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, Oxford, UK. Oxford Internet Institute.

Brauner, J. (2015). *Formalizations for Geooperators – Geoprocessing in Spatial Data Infrastructures*. PhD thesis, TU Dresden, Dresden, Germany.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Castellano, G., Fanelli, A., and Torsello, M. (2013). Web Usage Mining: Discovering Usage Patterns for Web Applications. In J. Velsquez, V. Palade, L. J., editor, *Advanced Techniques in Web Intelligence - 2. Studies in Computational Intelligence*, volume 452, pages 75–104. Springer, Berlin.

Choi, H. and Varian, H. (2012). Predicting the Present with Google Trends. *Economic Record*, 88(1):2–9.

Ferrara, E., De Meo, P., Fiumara, G., and Baumgartner, R. (2014). Web data extraction, applications and techniques: A survey. *Knowledge-Based Systems*, 70:301–323.

Gao, S. and Goodchild, M. F. (2013). Asking Spatial Questions to Identify GIS Functionality. In *Fourth International Conference on Computing for Geospatial Research and Application (COM. Geo)*, pages 106–110. IEEE.

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., and Watts, D. J. (2010). Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences of the United States of America*, 107(41):17486–17490.

Hecht, B., Hong, L., Suh, B., and Chi, E. H. (2011). Tweets from Justin Bieber's Heart: The Dynamics of the Location Field in User Profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 237–246, New York. ACM.

Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T., and Weitzner, D. (2008). Web Science: An Interdisciplinary Approach to Understanding the Web. *Communications of the ACM*, 51(7):60–69.

Hey, T., Tansley, S., Tolle, K. M., et al. (2009). *The Fourth Paradigm: Data-intensive Scientific Discovery*, volume 1. Microsoft Research, Redmond, WA.

Hinsen, K. (2014). Computational science: shifting the focus from tools to models. *F1000Research*, 3(101).

Hofer, B., Mäs, S., Brauner, J., and Bernard, L. (2017). Towards a knowledge base to support geoprocessing workflow development. *International Journal of Geographical Information Science*, 31(4):694–716.

Hollenstein, L. and Purves, R. (2010). Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science*, 1:21–48.

Hu, Y., Janowicz, K., Prasad, S., and Gao, S. (2015). Enabling semantic search and knowledge discovery for ArcGIS Online: A linked-data-driven approach. In *AGILE 2015*, pages 107–124. Springer, Berlin.

Janowicz, K., Van Harmelen, F., Hendler, J. A., and Hitzler, P. (2014). Why the data train needs semantic rails. *AI Magazine*, 36(1).

Kuhn, W. (2012). Core concepts of spatial information for transdisciplinary research. *International Journal of Geographical Information Science*, 26(12):2267–2276.

Kuhn, W. and Ballatore, A. (2015). Designing a Language for Spatial Computing. In Bacao, F., Santos, M. Y., and Painho, M., editors, *AGILE 2015: Geographic Information Science as an Enabler of Smarter Cities and Communities*, pages 309–326. Springer, Berlin.

Kveladze, I., Kraak, M.-J., and van Elzakker, C. P. (2013). A methodological framework for researching the usability of the space-time cube. *The Cartographic Journal*, 50(3):201–210.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. *Science*, 343(6176):1203–1205.

Lemmens, R. (2006). *Semantic interoperability of distributed geo-services.* PhD thesis, TU Delft, Delft, The Netherlands.

Ludäscher, B., Lin, K., Bowers, S., Jaeger-Frank, E., Brodaric, B., and Baru, C. (2006). Managing scientific data: From data integration to scientific workflows. *Geological Society of America Special Papers*, 397:109–129.

McKenzie, G., Janowicz, K., Gao, S., Yang, J.-A., and Hu, Y. (2015). POI pulse: A multi-granular, semantic signature–based information observatory for the interactive visualization of big geosocial data. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 50(2):71–85.

Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, pages 248–256. Association for Computational Linguistics.

Scheider, S. and Ballatore, A. (2018). Semantic typing of linked geoprocessing workflows. *International Journal of Digital Earth*, 11(1):113–138.

Scheider, S., Ostermann, F. O., and Adams, B. (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.

Stefanidis, A., Crooks, A., and Radzikowski, J. (2013). Harvesting ambient geospatial information from social media feeds. *GeoJournal*, 78(2):319–338.

Steiniger, S. and Hunter, A. J. (2013). The 2012 free and open source GIS software map - A guide to facilitate research, development, and adoption. *Computers, Environment and Urban Systems*, 39:136–150.

Stell, J. G. and Worboys, M. F. (1999). Generalizing Graphs Using Amalgamation and Selection. In Güting, R. H., Papadias, D., and Lochovsky, F., editors, *Advances in Spatial Databases: 6th International Symposium, SSD'99 Hong Kong, China, July 20—23, 1999 Proceedings*, pages 19–32. Springer, Berlin.

Stephens-Davidowitz, S. I. (2013). *Essays using Google Data*. PhD thesis, Harvard University, Cambridge, MA.