Annals of the American Association of Geographers

Routledge
Taylor & Francis Group

Where is the provenance? Ethical replicability and reproducibility in GIScience and its critical applications

Journal:	*Annals of the American Association of Geographers*
Manuscript ID	AN-2019-1342.R2
Manuscript Type:	Revised Manuscript
Manuscript Category:	Geographic Methods
Key Words:	GIScience, provenance, location privacy, data quality, disaster response

SCHOLARONE™
Manuscripts

Where is the provenance? Ethical replicability and reproducibility in GIScience and its critical applications

As replicability and reproducibility (R&R) crises develop within emerging convergent inquiry, ethical use of provenance information is central to the establishment and preservation of trust in critical applications of GIScience and geospatial technologies. Today large volumes of geospatial data are generated at high velocity from satellite sensors and unmanned aircraft systems, citizen sensors, geolocation-based data services, global navigation satellite systems, etc. The extensive use of these data for applications such as disaster and humanitarian response raises the issue of R&R from competing perspectives of location privacy and geospatial data quality. While geospatial data can be integrated and linked with contextual information to identify individuals' movements, steps taken to ensure privacy can complicate the multiuser development of high-quality geospatial workflows. Provenance information as digital records of historical (retrospective) and potential future (prospective) geospatial processes is often overlooked, misunderstood, or inadequately addressed. We explore the relationship between provenance information, location privacy, and geospatial data quality in the context of R&R with a focus on disaster analytics. We argue that in the era of big data and deep learning, GIScientists and associated institutions bear greater responsibility both for geospatial workflow quality and for location privacy. Given vastly heterogenous computational landscapes, we provide practical recommendations for ethically driven provenance and R&R research and development within the GIScience community and beyond.

Keywords: GIScience; provenance; location privacy; data quality; disaster response

For Peer Review Only

Introduction and statement of the problem

A "culture of convergence" allows innovative problem solving not only for transdisciplinary researchers but all relevant stakeholders (e.g., farmers, drone technology providers, healthcare professionals, etc.) to improve quality of life (National Academies 2019a). Convergence research was one of "10 big ideas" identified in 2016 by National Science Foundation (2020): "From its inception, the convergence paradigm intentionally brings together intellectually-diverse researchers to *develop effective ways of communicating across disciplines by adopting common frameworks and a new scientific language*, which may, in turn, afford solving the problem that engendered the collaboration, developing novel ways of framing research questions, and opening new research vistas."

Geography is a key participant, and GIScience has the potential to become a major "clearinghouse" in convergent inquiry given its long history of convergence research to solve real-world problems (Center for Advanced Spatial Technologies 2019; University Consortium for Geographic Information Science 2020, Goodchild and Haining 2004). What dedicated efforts or enabling technologies can best help GIScience to realize this potential? The answer is important for the extended health of our discipline, for trust in its critical applications, and for how well it addresses location privacy, intellectual property, and other individual stakeholder interests. Clearly, a meaningful part of the answer has to do with sources, ownership, history, and reproducibility and replicability (R&R; National Academies 2019b) of the information being distributed as GIScience.

This paper highlights three unique but interrelated concepts that contribute to R&R success in a convergent GIScience paradigm: 1) provenance, 2) location privacy, and 3) data quality. By presenting these concepts in the context of disaster response, we

For Peer Review Only

attempt to provide recommendations for R&R-related practices that are relevant to other critical geospatial applications (e.g., agriculture and food security) and convergent stakeholder interests (e.g., intellectual property).

The paper is organized as 1) a discussion of the concepts of provenance, location privacy and data quality; 2) a discussion of current practices to address provenance, location privacy and data quality concerns in R&R-enabled GIScience research; 3) an illustration of the implications of failing to adhere to the afore-mentioned concepts in the context of disaster response where convergence research is predominant; and 4) concluding remarks and a few recommendations from researcher perspectives.

Concepts affecting reproducibility and replicability within convergent GIScience

Provenance

There is an abundance of geospatial and other data of multiple varieties, and yet an underwhelming ability via current public methodologies to ascertain their provenance (sources, ownership, and history). In a study on research integrity, the National Academies (2017) describe evidence of the R&R crisis in scientific publications, including the number of retractions due to intentional fraud. However, of greater interest to GIScience is their note that "widespread utilization of information technologies in research may… introduce new sources of *unintentional* error and irreproducibility of results" (National Academies 2017, 38; emphasis added). Geographers and others involved in convergent GIScience should consider the implications of this observation given the fast pace of digital evolution that by its very nature introduces an unintended versioning crisis. From computer systems, Buneman (2013) describes a lack of appreciation for the machine-managed provenance found in

For Peer Review Only

everyday software version control systems (e.g., Git-based repositories used frequently in GIS and increasingly in other-than-code applications). The systematic treatment of this provenance, he argues, helps to prevent software engineering from completely falling apart. While R&R challenges (National Academies 2019b) associated with geospatial versioning crises appear symptomatic of deeper communication issues, the question of provenance information (our focus here) seems closer to the core of GIScience.

Dating back more than 700 years in Old French, *provenance* derives from the Latin provenire, and within the art domain "entails an artifact's complete ownership history, but ideally will also include artistic, social, and political influences upon the work from its creation to the present day" (Tullis et al. 2015, 403). Creation of such trusted provenance information is a formal process that often garners significant investment when used for artwork authentication (IFAR 2013; Yeide et al 2001). Fast forwarding to provenance in computer systems, it is remarkable to note geography's early influence. From Nicholas Chrisman's (1983) treatment of lineage in a GIS lifecycle to David Lanter's (1989) dissertation on geodatabase lineage tracing and subsequent patent (1993), geography has been a pioneering actor in provenance at various levels of digital granularity or detail. (Finely granulated provenance, for example, might include source information such as individual satellite image pixel values, and coarse granularity might only provide a reference to an archival copy of the relevant satellite image.) Tullis et al. (2015) examine provenance in a GIScience context and cite the Worldwide Web Consortium (W3C) Provenance Incubator Group's working definition (Gil et al. 2010): "Provenance of a resource is a record that describes entities and processes involved in producing and delivering or otherwise influencing that resource. Provenance provides a critical foundation for assessing authenticity,

For Peer Review Only

enabling trust, and allowing reproducibility. Provenance assertions are a form of contextual metadata and can themselves become important records with their own provenance."

One-to-one cardinality between data and descriptive metadata (where provenance is often relegated in GIS; e.g., lineage information as intended by ISO 2019) is different than one-to-many cardinality between data and *contextual* metadata (where provenance is more accurately characterized; Gil et al. 2010). According to the AIMS Team of the FAO (2019), contextual metadata "concerns persons, organisations, projects, funding, outputs, facilities and equipment…to assess the relevance and quality of an object (dataset, software) for their current purpose" (emphasis removed). The difference between these two types of metadata is critical and has possibly led to much confusion (Tullis et al. 2019; Tullis et al. 2015; Gil et al. 2010).

As with metadata, multiple forms and definitions of provenance are found in computer systems where distinctions are made, for example, between file versus database management system (DBMS) environments (Moreau 2010). Buneman's (2013, 11) suggested "change of attitude" encourages us to "worry less about what provenance is and concentrate more on what we can do with it once we have it." The existence of detailed geospatial provenance records, including sources, ownership, and history, naturally raises questions of privacy and data quality. This situation has significant societal and policy implications during disaster events when both static and dynamic data sets available via crowdsourcing to remote sensing are used to aid with decision-making efforts (Qadir et al. 2016; Naja et al. 2010).

Location privacy

Personal privacy, or "the right to be alone" (Warren and Brandeis 1890), has a tradition

For Peer Review Only

of protection in the US and other countries. By contrast, "location privacy" is an emerging though somewhat nebulous geospatial concept. Location privacy refers to an individual's expectation to move freely in geographic public space without the fear of being recorded, and of those records being used by third parties (Blumberg and Eckersley 2009; Kar, Crowsey and Zale 2012). Several bills, such as the Online Communications and Geolocation Protection Act of 2013, the California Consumer Privacy Act of 2018, and the Geolocation Privacy and Surveillance or GPS Act of 2011 aim to prevent government entities from obtaining or disclosing users' location information without consent and using the information for any purpose, and require commercial entities to protect the online privacy of users and their data (National Coordination Office for Space-Based Positioning, Navigation, and Timing 2019; National Conference on State Legislatures 2019). While laws have been enacted in a fragmented fashion to protect online privacy, user data, and access to geolocation information, there is no comprehensive law comparable to the Privacy Act of 1974 to protect location privacy in the era of big data in the US.

With interconnected devices, it is increasingly possible to breach personal privacy using location data when there is no restriction on how location data are being used and by whom. For instance, a location service from Securus Technologies was reported to have allowed a former sheriff to repeatedly track people's location in real time using location data purchased from third party telecoms (Valentino-DeVries 2018). Undeniably, individual location information available from social media data, location-based services and mobile phones enable search and rescue operations during disasters as well as for contact tracing during the COVID-19 pandemic to reduce infections. However, current technologies and analytics are still not equipped to address the privacy concerns and legal implications of using these data sets during disaster response

For Peer Review Only

(National Research Council 2013; Soltani et al. 2020). Such issues force us to ask to what extent GIScience and convergent GIScience efforts protect privacy while enabling innovative applications of location-based services.

Data quality

Geospatial data quality depends on the purpose for which data are used, the cost and value of the data to customers, and the processes by which data are generated (Karr et al. 2001). Despite best efforts within a collaborative team, there may be confusion about who is responsible for determining data quality for intended purposes. As Chrisman (1986, 352) suggests, "evaluation and judgment of fitness of use must be the responsibility of the user, not the producer. To carry out this responsibility, the user must be presented with much more information to permit an informed decision."

Given the active history of geospatial data, standardized methods and frameworks have been developed to assess, describe, and quantify geospatial data quality. Moellering et al. (1988) identified five components to assess geospatial data quality: 1) lineage (comparable to provenance), 2) positional accuracy, 3) attribute accuracy, 4) logical consistency, and 5) completeness. After the Federal Geographic Data Committee's 1994 metadata standards were developed, ongoing ISO Technical Committee efforts have generated several geospatial data quality standards. ISO 8402, for example, addresses quality management and quality assurance, and ISO 19157 focuses on the quality of geospatial data. Interestingly, the latest iteration of ISO 19115 "Geographic Information – Metadata - Part 2" standard, has clarified its focus away from data structures to acquisition and processing (ISO 2019; Tullis et al. 2019). These standards use quantitative (e.g., spatiotemporal uncertainty, thematic accuracy, resolution, consistency, and completeness) and qualitative measures (e.g., purpose,

For Peer Review Only

usage, and lineage) to assess data quality (Veregin 1999). However, in the absence of cohesive metadata and provenance associated with geospatial data, it is difficult to assess data quality (Moroni et al. 2019), a clear understanding of which is vital to support R&R success.

Data produced from different organizations, generated by different devices (citizen and remote sensors as well as mobile phones) in different formats are integrated to aid with disaster response activities. For instance, flood extent output determined by hydrologic models and earth observation data are used in conjunction with built environment data to determine damaged structures or identify shelter locations or evacuation points. Although databases such as the global level emergency disaster database EM-DAT (Centre for Research on the Epidemiology of Disasters 2020) provide easy access to some of these data sets, the variable formats along with inconsistent standards makes using these data effectively during disasters difficult. There is a need to assess data quality and make these data available in interoperable formats for easy integration as well as R&R of workflows during disaster response (Fakhruddin, Chu, and Li 2019; OGC 2018).

For Peer Review Only

Current practice and recommendations for innovation

Provenance

As the world geospatial market is poised to exceed approximately $88 billion in 2020 (Narain 2018), a major disruption in geospatial data collection is occurring through activities based on unmanned aircraft system (UAS; Tullis et al. 2019). Whereas data collection has historically required sophisticated and perhaps trusted team experience (such as NASA), there is now a trend toward “personal remote sensing” (Jensen 2017). The concept of provenance is insufficiently explored in GIScience and remote sensing,

and there is certainly no widely accepted standard in place. UAS manufacturers, for example, generate proprietary binary logfiles to capture the history of a drone's GNSS positions, gimbal orientations, etc. Depending on the level of integration between platform and sensor(s), a portion of this information may be automatically transferrable to image-stream metadata and subsequently to copies of the data. Geoprocessing environments often capture workflow-level provenance through similarly proprietary logfile formats.

Proprietary graphic block programming environments (e.g., Esri's ModelBuilder) represent a visually augmented form of code that may contain both valuable retrospective (historical software execution) and prospective (potential future) provenance information (Tullis et al. 2015). While original geoprocessing steps are often preserved and possibly curated (e.g., WorldView-3 satellite images, UAS image streams), intermediate datasets within complex workflows are often discarded or effectively lost in archives due to storage and management costs. Over time, human experts with knowledge of specific geospatial provenance (e.g., how to interpret vague documentation in Python scripts) change jobs, retire, etc. and non-backwards compatible software versions march forward. In short, while there is ample geospatial provenance information in many forms, provenance records are rarely interoperable.

There are at least three avenues to mitigate the current chaotic provenance scenario. First, GIScience should encourage more practical demonstrations of multiuser provenance interchange that accelerate R&R, regardless of whether the software environments are open-source or commercial, workstation or cloud based, or code versus graphic block programming (Tullis et al. 2015, 2019). Much work remains in how to compress, store, and otherwise manage provenance records throughout their lifecycle (Sun et al. 2020), issues that are continually emerging in discussions on

For Peer Review Only

GIScience workflows (OGC 2020). Even though the Worldwide Web Consortium (W3C) created the powerful PROV data model (PROV-DM; Moreau and Missier 2013) almost a decade ago, GIScience applications of PROV have been very rare (Tullis et al. 2019). Successful, practical demonstrations of AI and hybrid GIScience workflow management (e.g., Sun et al. 2020) challenge software engineering teams to accelerate support for provenance interchange and workflow standards and specifications from authoritative sources such as W3C and OGC.

Second, a convergent GIScience initiative should seek to balance practical aspects (e.g., the mechanics of how provenance records may be interchanged) with a systematic study of how a provenance foundation can support R&R, data quality, and other benefits (Gil et al. 2010). This can include systematic analysis of both retrospective provenance information and simulation of prospective software execution (by hypothetical individuals) as it relates to complex issues, such as location privacy.

Finally, open and commercial software innovators, industry investors, and policy stakeholders should be invited to participate in the above convergent activities and to consider the development of *geospatial privacy services* and other provenance-enabled web services (e.g., *data quality*, *intellectual property*, *export control*, etc.). Interestingly, Code Ocean (2020) has demonstrated remarkable R&R services using code-based source control and this platform addresses privacy through repository access constraints. However, to be able to leverage both code and other-than-code provenance records in a potentially unlimited variety of commercial and open formats, such services will very likely depend upon the active adoption of provenance standards such as possible future geospatial extensions of PROV. As will be discussed below, domain-specific practices and requirements should be considered in the development of provenance-enabled services.

For Peer Review Only

Location privacy

Individuals knowingly and unknowingly are producing large volumes of geospatial data through multiple channels. Examples include (i) location-based services that require users to share their identities and location information; (ii) geo-social or proximity-sensitive services like WeChat and Facebook; (ii) web mapping and information services like Yelp and Google Maps; (iii) healthcare systems like Fitbit; and (iv) augmented reality games like PokemonGo (Liu et al. 2018). Location data generated by these services are used for myriad applications in real- and near-real time, alone or in combination with other data sources. For example, during Hurricanes Harvey in 2017 and Matthew in 2016, social media data were integrated with imagery obtained from UAS to aid in response (Cox 2016; Wright 2017).

While perhaps less than 3% of social media data contain location information that may be useful in situations like disaster response (Leetaru et al. 2013), advancements in GeoAI and deep learning have made it possible, for example, to geotag a large proportion of tweets (Minami and Borisov 2010; Do et al. 2017). These deep learners cross-link information across databases to identify an individual's demography, land-use zones, movement patterns, and photos to subsequently predict location information (Wayand, Kostrikov and Philbin 2016; Ren et al. 2017; Li et al. 2018). Location data can also be linked with other information to create a profile of an individual's daily activities and pattern of life (Liu et al. 2018) as is seen in the case of COVID-19 spread in Wuhan, China (Wu et al. 2020).

To protect location privacy and prohibit extraction of personal information from location data, several techniques are used. These techniques include (a) cryptography (using encryption to protect user position); (b) anonymity (removing links between location and identity information); (c) obfuscation (reducing the precision of location

For Peer Review Only

information); and (d) caching mechanisms (storing points of interest in a cache separate from location information) (Liu et al. 2018). Differential privacy approaches, such as Laplace mechanism, are used to fudge location data such that the output protects privacy (Andrés et al. 2013). However, adversarial techniques, such as Markov chain models and deep learners that combine contextual information with location information, are used to reduce the effectiveness of privacy protection techniques (Minami and Borisov 2010; Li et al. 2018). For R&R-enabled GIScience to address privacy, it may be more important to understand *how* the location information is being used rather than *whether* it could be obtained. That question can best be answered through a greater focus on developing theory and applications surrounding retrospective provenance records; by definition, provenance is a key for how data have been used in each scientific workflow (National Academies 2019b).

Data quality

Inherent biases in data collection and implicit biases in datasets, issues frequently raised in the realm of big data and machine learning, also apply to geospatial data. Such biases stem from many sources. For example, selection and demographic biases are associated with citizen-derived data. Social media data are often produced by young adults with smart phones (Pew Research Center 2018). Hence, citizen-derived geospatial big data tend not to represent the entirety of the population. Haklay (2010) reported that the volunteer-based OpenStreetMap data were detailed and abundant for rich (but not for poor) neighborhoods. Thus, determining completeness and biases in these data may be extremely difficult. Unlike formal survey data, the internal and external validity of geospatial big data, therefore, are often undermined.

For Peer Review Only

Research is dependent on data quality which can propagate through analysis and modelling, and therefore, problematic issues need to be identified, addressed, and possibly ameliorated. Standardized information about such issues and quality concerns should be linked with datasets, research results, and potentially real-time applications. While international standards are in place to assess the quality of geospatial data, no such standard exists to assess the quality of geospatial big data (those obtained from urban and citizen sensors). The United Nations Economic Commission for Europe (UNECE Big Data Quality Task Team 2014) presented a framework to assess big data quality across three hyper dimensions and 13 quality dimensions (Table 1). While OGC (2018) has provided guidelines to produce datasets in specific standards for disaster resilience, how and when to use these dimensions to ensure data quality in an R&R paradigm falls on the user community. However, provenance is a major key to organizing, linking, and prioritizing these and other quality dimensions not presently considered.

Table 1. Framework to assess big data quality across three hyper dimensions and 13 quality dimensions. Adapted from UNECE Big Data Quality Task Team (2014).

Hyperdimension	**Quality dimension**
Source	Institutional/business environment
	Privacy and security
Metadata	Complexity
	Completeness
	Usability
	Time-related factors
	Linkability
	Coherence – consistency
	Validity
Data	Accuracy and selectivity
	Linkability
	Coherence – consistency
	Validity

For Peer Review Only

Illustration in disaster analytics

Geospatial data and tools are extensively used across emergency management phases (National Research Council 2007, Kevany 2008). With the availability of geospatial big data before, during, and after a hazard event at fine temporal resolution, these data are used in response and recovery efforts. For instance, tweets available within two minutes after the first onset of tremors during the 2011 Tohuku earthquake provided information about the earthquake, the tsunami, and the Fukushima nuclear disaster (Pu & Kitsuregawa 2013). During 2012 Superstorm Sandy, Foursquare (a location-based social network site) aided with search and rescue operations by determining the spatial distribution of visitors before and after the hurricane (Pu & Kitsuregawa 2013). As exemplified in the COVID-19 pandemic and the 2010 Haiti cholera epidemic, social media enables information dissemination and social connectivity, and data from mobile phones and WeChat help determine disease spread based on mobility (Lomas 2020; Wu et al. 2020; Merchant and Lurie 2020; Bengtsson et al. 2015).

Despite obvious usability, geospatial big data often do not contain the kinds of information that might make them relevant for response activities (Liu et al. 2019). For instance, sensors and cameras deployed in a city provide geospatial data to help with traffic congestion, thwart crime or terrorism, and respond to disasters. However, these datasets are generated in non-standardized formats according to manufacturer requirements and service provider needs (Voigt and Bright 2016). Due to openness of these data, third-party developers use these data to develop situational-awareness tools for emergency response (Wendling, Radisch and Jacobzone 2013). For instance, several web maps have been developed to visualize the outbreak of COVID-19 (The New York Times 2020; Information Is Beautiful 2020). Mobility data derived from mobile phones, aggregated to different spatial scales of analysis, have been critical to understand the

For Peer Review Only

spread of the COVID-19 pandemic. The aggregate mobility data such as Google mobility reports (Google 2020) rarely provide any information about the quality and provenance of input data (such as, what percent of sample data used to generate mobility reports at country level) or the methods implemented to ensure location privacy of users.

While end results are informative, data quality is questionable and location privacy is ignored as individuals have limited-to-no choice or ability to consent to how their data are being used. Not having answers to these questions also influences how the data are being used for decision-making. Mobility patterns have been central to making decisions on shelter-in-place policies in the United States (Passy 2020). With reduction in COVID-19 cases, shelter-in-place policies have mostly been lifted, thereby leading to higher mobility and increase in infected cases (Bella 2020). As pointed out by Ienca and Vayena (2020), in case of COVID-19 and other disasters, it is crucial to ensure privacy while using location, quality, and provenance information. Going forward, it is critical to address these concepts while using geospatial data sets and GIScience theories in convergence research to aid with disaster response.

For Peer Review Only

Discussion and conclusion

In GIScience research, R&R concepts (Association for Computing Machinery 2018) require that the same results be obtained under the same or different experimental setups. When there is the possibility of losing location privacy, countervailing techniques (discussed above) can be implemented but only if such possibilities are first explored, understood, and acknowledged by stakeholders. A lack of clear provenance information linked to original and derived geospatial data (including that which may be privacy-sensitive) can produce different results for even the same experimental setup

(and location), thereby reducing R&R. In the case of temporal studies, even if provenance is maintained, privacy assurance may prevent its free interchange and thus similarly reduce R&R. Likewise, variation in data quality over space or time can reduce R&R in GIScience. As discussed through case studies in Kitzes et al. (2018), R&R requires (i) maintaining provenance, data quality assessment, privacy protection techniques, workflow implementation, and software and analytical techniques, and (ii) ensuring access to the same data and sources across space and time.

Currently, researchers are mostly left on their own to navigate the complex terrain of geospatial provenance, location privacy, and data quality. We think this individually defined approach is ill-suited for the scope and magnitude of potential consequences in critical applications of GIScience. Yet, the growing number of formalized efforts to translate responsibility into codes of ethics or best practices are fragmented across disparate professional, academic, and commercial spheres. Location privacy is murkier due to the use of publicly available geospatial big data and in the case of applied research, near-term good may be countered by longer term harms.

With geospatial provenance currently in the periphery, systematic treatment of this important *contextual* metadata is integral to any antidote for mitigating confusion in geospatial big data privacy and quality. Such an approach, including grassroots practical demonstrations and a convergent GIScience focus, requires strong ethics. There are always risks of unintended consequences, including the potential for increased ability to interchange adversarial or unethical methods. Yet, the question of *where* to find geospatial provenance information will be paramount in allowing researchers and other stakeholders to better understand geospatial quality, location privacy, and other benefits for critical GIScience applications.

For Peer Review Only

References

AIMS Team of the FAO. 2019. Metadata for description, discovery and contextualisation. http://aims.fao.org/activity/blog/metadata-description-discovery-contextualisation-check-rda-metadata-catalog (last accessed 10 June 2020).

Andrés, M. E., N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi. 2013. Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, CCS '13., 901–914. Berlin, Germany: Association for Computing Machinery https://doi.org/10.1145/2508859.2516735 (last accessed 1 April 2020).

Association for Computing Machinery. 2018. Artifact review and badging. https://www.acm.org/publications/policies/artifact-review-badging (last accessed 30 April 2019).

Bella, T. 2020. Places without social distancing have 35 times more potential coronavirus spread, study finds. *Washington Post* 15 May. https://www.washingtonpost.com/nation/2020/05/15/social-distancing-study-coronavirus-spread/ (last accessed 12 June 2020).

Bengtsson, L. et al. 2015. Using mobile phone data to predict the spatial spread of cholera. *Scientific Reports* 5 (1):1–5.

Blumberg, A. J., and P. Eckersley. 2009. On locational privacy, and how to avoid losing it forever. *Electronic Frontier Foundation* :8.

Buneman, P. 2013. The providence of provenance. In *Big Data*, eds. G. Gottlob et al., 7–12. Berlin, Germany: Springer http://link.springer.com/10.1007/978-3-642-39467-6_3 (last accessed 27 March 2015).

For Peer Review Only

Center for Advanced Spatial Technologies. 2020. *Convergent GIScience priorities*. Fayetteville, AR. https://cast.uark.edu/events/giscience.php (last accessed 9 November 2019).

Centre for Research on the Epidemiology of Disasters. 2020. EM-DAT: The international disasters database. https://www.emdat.be/ (last accessed 10 June 2020).

Chrisman, N. R. 1986. Obtaining information on quality of digital data. In *AutoCarto Proceedings of the International Symposium on Computer-Assisted Cartography*, 350–358. London, England: Cartography and Geographic Information Society.

———. 1983. The role of quality information in the long-term functioning of a geographic information system. *Automated Cartography* (6):302–312.

Code Ocean. 2020. Code Ocean. https://codeocean.com/ (last accessed 12 June 2020).

Cox, J. 2016. How Twitter and a drone saved a man and his dog from Hurricane Matthew. *Vice*. https://www.vice.com/en_us/article/9a3gmd/how-twitter-and-a-drone-saved-a-man-and-his-dog-from-hurricane-matthew (last accessed 15 November 2019).

Do, T. H., D. M. Nguyen, E. Tsiligianni, B. Cornelis, et al. 2017. Multiview deep learning for predicting Twitter users' location. *arXiv:1712.08091 [cs.LG]* :12.

Fakhruddin, B., E. Chu, and G. Li. 2019. *Next generation disaster data infrastructure - White paper*. Paris: CODATA Task Group, Linked Open Data for Global Disaster Risk Research (LODGD). http://www.irdrinternational.org/2019/09/17/next-generation-disaster-data-infrastructure/.

Gil, Y. et al. eds. 2010. The foundations for provenance on the web. *Foundations and Trends in Web Science* 2 (2–3):99–241.

For Peer Review Only

Goodchild, M. F., and R. P. Haining. 2004. GIS and spatial data analysis: Converging perspectives. *Papers in Regional Science* 83 (1):363–385.

Google. 2020. COVID-19 community mobility reports. https://www.google.com/covid19/mobility?hl=en (last accessed 12 June 2020).

Haklay, M. 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design* 37 (4):682–703.

Ienca, M., and E. Vayena. 2020. On the responsible use of digital data to tackle the COVID-19 pandemic. *Nature Medicine* :1–2.

IFAR. 2013. Provenance guide. http://www.ifar.org/provenance_guide.php.

Information is Beautiful. 2020. Coronavirus infection trajectories. https://informationisbeautiful.net/visualizations/covid-19-coronavirus-infographic-datapack/ (last accessed 1 April 2020).

ISO. 2019. ISO 19115-2:2019 Geographic information - Metadata - Part 2: Extensions for acquisition and processing. *ISO*. http://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/70/67039.html (last accessed 20 May 2019).

Jensen, J. R. 2017. *Drone aerial photography and videography: Data collection and image interpretation* 1.2. Apple. https://itunes.apple.com/us/book/drone-aerial-photography-and-videography/id1283582147?mt=11 (last accessed 16 February 2018).

Kar, B., R. C. Crowsey, and J. J. Zale. 2012. The myth of location privacy in the United States: Surveyed attitude versus current practices. *The Professional Geographer* 65 (1):47–64.

For Peer Review Only

Karr, A. F., A. P. Sanil, and J. Sacks. 2001. *Affiliates workshop on data quality*. Research Triangle Park, NC: National Institute of Statistical Sciences.

Kevany, M. J. 2008. Improving geospatial information in disaster management through action on lessons learned from major events. In *Geospatial Information Technology for Emergency Response*, International Society for Photogrammetry and Remote Sensing (ISPRS) Book Series., eds. S. Zlatanova and J. Li, 3–19. AK Leiden, The Netherlands: Taylor & Francis/Balkema.

Kitzes, J., D. Turek, and F. Deniz eds. 2018. *The practice of reproducible research: Case studies and lessons from the data-intensive sciences*. Oakland, CA: University of California Press. http://www.practicereproducibleresearch.org.

Lanter, D. P. 1993. Method and means for lineage tracing of a spatial information processing and database system. :87.

———. 1989. Techniques and method of spatial database lineage tracing.

Leetaru, K., S. Wang, G. Cao, A. Padmanabhan, et al. 2013. Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday* 18 (5). https://journals.uic.edu/ojs/index.php/fm/article/view/4366 (last accessed 13 November 2019).

Li, H., H. Zhu, S. Du, X. Liang, et al. 2018. Privacy leakage of location sharing in mobile social networks: attacks and defense. *IEEE Transactions on Dependable and Secure Computing* 15 (4):646–660.

Liu, B., W. Zhou, T. Zhu, L. Gao, et al. 2018. Location privacy and its applications: A systematic study. *IEEE Access* 6:17606–17624.

Liu, X., B. Kar, C. Zhang, and D. M. Cochran. 2019. Assessing relevance of Tweets for risk communication. *International Journal of Digital Earth* 12 (7):781–801.

For Peer Review Only

Lomas, N. 2020. Germany's COVID-19 contacts tracing app to link to labs for test result notification. *TechCrunch*. https://social.techcrunch.com/2020/04/23/germanys-covid-19-contacts-tracing-app-to-link-to-labs-for-test-result-notification/ (last accessed 12 June 2020).

Merchant, R. M., and N. Lurie. 2020. Social media and emergency preparedness in response to novel coronavirus. *JAMA*.

Minami, K., and N. Borisov. 2010. Protecting location privacy against inference attacks. In *Proceedings of the 17th ACM conference on Computer and communications security - CCS '10*, 711. Chicago, Illinois, USA: ACM Press http://portal.acm.org/citation.cfm?doid=1866307.1866406 (last accessed 13 November 2019).

Moellering, H. et al. 1988. The proposed standard for digital cartographic data. *The American Cartographer* 15 (1):9–140.

Moreau, L. 2010. The foundations for provenance on the web. *Foundations and Trends in Web Science* 2 (2–3):99–241.

Moreau, L., and P. Missier eds. 2013. PROV-DM: The PROV data model. http://www.w3.org/TR/2013/REC-prov-dm-20130430/#section-example-two (last accessed 7 August 2014).

Moroni, D. F. et al. 2019. *Understanding the various perspectives of Earth science observational data uncertainty*. ESIP. https://esip.figshare.com/articles/Understanding_the_Various_Perspectives_of_Earth_Science_Observational_Data_Uncertainty/10271450 (last accessed 1 April 2020).

For Peer Review Only

Naja, I., L. Moreau, and A. Rogers. 2010. Provenance of decisions in emergency response environments. In *Provenance and annotation of data and processes*, Lecture Notes in Computer Science., eds. D. L. McGuinness, J. R. Michaelis, and L. Moreau, 221–230. Berlin, Heidelberg: Springer Berlin Heidelberg http://link.springer.com/10.1007/978-3-642-17819-1_25 (last accessed 12 June 2020).

Narain, A. 2018. GIS and spatial analytics market to touch US$ 88.3 billion by 2020! *Geospatial World*. https://www.geospatialworld.net/blogs/gis-and-spatial-analytics-market/ (last accessed 12 November 2019).

National Academies of Sciences, Engineering, and Medicine. 2017. *Fostering integrity in research*. Washington, D.C.: National Academies Press. https://www.nap.edu/catalog/21896/fostering-integrity-in-research (last accessed 12 April 2017).

———. 2019a. *Fostering the culture of convergence in research: Proceedings of a workshop*. Washington, D.C.: The National Academies Press. https://www.nap.edu/catalog/25271/fostering-the-culture-of-convergence-in-research-proceedings-of-a (last accessed 9 August 2019).

———. 2019b. *Reproducibility and replicability in science*. Washington, D.C.: National Academies Press. https://www.nap.edu/catalog/25303 (last accessed 17 May 2019).

National Conference on State Legislatures. 2019. Consumer data privacy legislation. http://www.ncsl.org/research/telecommunications-and-information-technology/consumer-data-privacy.aspx (last accessed 15 November 2019).

For Peer Review Only

National Coordination Office for Space-Based Positioning, Navigation, and Timing. 2019. GPS: The Global Positioning System. https://www.gps.gov/ (last accessed 10 November 2019).

National Research Council. 2013. *Public response to alerts and warnings using social media: Report of a workshop on current knowledge and research gaps*. Washington, D.C.: National Academies Press. http://www.nap.edu/catalog/15853 (last accessed 12 June 2020).

———. 2007. *Successful response starts with a map: Improving geospatial support for disaster management*. Washington, DC: The National Academies Press. http://www.nap.edu/catalog/11793 (last accessed 12 June 2020).

National Science Foundation. 2020. What is convergence? https://www.nsf.gov/od/oia/convergence/index.jsp (last accessed 10 June 2020).

OGC. 2018. *OGC Development of Disaster Spatial Data Infrastructures for Disaster Resilience*. Open Geospatial Consortium. https://portal.opengeospatial.org/files/18-087r5.

———. 2020. Workflow DWG. http://www.opengeospatial.org/projects/groups/workflowdwg (last accessed 21 May 2019).

Passy, J. 2020. Coronavirus cases plummet by 44% due to shelter-in-place orders, according to this study drawing on CDC data. *MarketWatch*. https://www.marketwatch.com/story/coronavirus-cases-drop-by-up-to-44-due-to-shelter-in-place-orders-study-drawing-on-cdc-data-shows-2020-05-04 (last accessed 12 June 2020).

For Peer Review Only

Pew Research Center. 2018. *Social media use in 2018*. Pew Research Center. https://www.pewresearch.org/internet/2018/03/01/social-media-use-in-2018/ (last accessed 12 June 2020).

Pu, C., and M. Kitsuregawa. 2013. Big data and disaster management: A report from the JST/NSF joint workshop. :28.

Qadir, J. et al. 2016. Crisis analytics: big data-driven crisis response. *Journal of International Humanitarian Action* 1 (1):12.

Ren, S., K. He, R. Girshick, and J. Sun. 2017. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (6):1137–1149.

Soltani, A., R. Calo, and C. Bergstrom. 2020. Contact-tracing apps are not a solution to the COVID-19 crisis. *Brookings*. https://www.brookings.edu/techstream/inaccurate-and-insecure-why-contact-tracing-apps-could-be-a-disaster/ (last accessed 12 June 2020).

Sun, Z., L. Di, A. Burgess, J. A. Tullis, et al. 2020. Geoweaver: Advanced cyberinfrastructure for managing hybrid geoscientific AI workflows. *ISPRS International Journal of Geo-Information* 9 (2):119.

The New York Times. 2020. Coronavirus map: tracking the global outbreak. *The New York Times* 1 April. https://www.nytimes.com/interactive/2020/world/coronavirus-maps.html (last accessed 1 April 2020).

Tullis, J. A. et al. 2015. Geoprocessing, workflows, and provenance. In *Remotely Sensed Data Characterization, Classification, and Accuracies*, Remote Sensing Handbook., ed. P. S. Thenkabail, 401–421. Boca Raton, FL: CRC Press.

For Peer Review Only

Tullis, J. A., K. Corcoran, R. Ham, B. Kar, et al. 2019. Multiuser concepts and workflow replicability in sUAS applications. In *Applications of Small Unmanned Aircraft Systems*, ed. J. B. Sharma, 35–56. New York, NY: CRC Press https://www.crcpress.com/Applications-of-Small-Unmanned-Aircraft-Systems-Best-Practices-and-Case/Sharma/p/book/9780367199241.

UNECE Big Data Quality Task Team. 2014. *A suggested framework for the quality of big data*. https://www.google.com/url?client=internal-element-cse&cx=015747183538840128847:7r0qtadycj4&q=https://statswiki.unece.org/download/attachments/108102944/Big%2520Data%2520Quality%2520Framework%2520-%2520final-%2520Jan08-2015.pdf%3Fversion%3D1%26modificationDate%3D1420725063663%26api%3Dv2&sa=U&ved=2ahUKEwiNiO-aufzpAhUCJKwKHQwBBSgQFjAAegQIARAC&usg=AOvVaw00jbcWQSHrWey4XiFCETFC.

University Consortium for Geographic Information Science. 2020. Initiative on convergence research of GIScience, humanities, and social sciences. https://www.ucgis.org/index.php?option=com_content&view=article&id=232:initiative-on-convergence-research&catid=38:committees (last accessed 11 June 2020).

Valentino-DeVries, J. 2018. Service meant to monitor inmates' calls could track you, too. *The New York Times* 10 May. https://www.nytimes.com/2018/05/10/technology/cellphone-tracking-law-enforcement.html (last accessed 12 November 2019).

Veregin, H. 1999. Data quality parameters. In *Geographical information systems*, eds. P. A. Longley et al., 177–189.

For Peer Review Only

Voigt, C., and J. Bright. 2016. The lightweight smart city and biases in repurposed big data. :7.

Warren, S. D., and L. D. Brandeis. 1890. The right to privacy. *Harvard Law Review* 4 (5):193.

Waterfield, B. 2011. Tom Tom sold driver's GPS details to be used by police for speed traps. 28 April. https://www.telegraph.co.uk/technology/news/8480702/Tom-Tom-sold-drivers-GPS-details-to-be-used-by-police-for-speed-traps.html#:~:targetText=Tom%20Tom%2C%20the%20satellite%20navigation,roadside%20speed%20traps%20and%20cameras. (last accessed 15 November 2019).

Wendling, C., J. Radisch, and S. Jacobzone. 2013. *The use of social media in risk and crisis communication*. https://www.oecd-ilibrary.org/governance/the-use-of-social-media-in-risk-and-crisis-communication_5k3v01fskp9s-en (last accessed 13 November 2019).

Weyand, T., I. Kostrikov, and J. Philbin. 2016. PlaNet - Photo geolocation with convolutional neural networks. In *Computer Vision – ECCV 2016*, eds. B. Leibe et al., 37–55. Cham: Springer International Publishing http://link.springer.com/10.1007/978-3-319-46484-8_3 (last accessed 13 November 2019).

Wright, T. 2017. By pitching in to help storm victims, drones earn new respect. *Air & Space Magazine*. https://www.airspacemag.com/daily-planet/after-two-hurricanes-drones-prove-useful-pinch-180964908/ (last accessed 15 November 2019).

Wu, J. T., K. Leung, and G. M. Leung. 2020. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *Lancet (London, England)* 395 (10225):689–697.

For Peer Review Only

Yeide, N. H., K. Akinsha, and A. L. Walsh. 2001. *The AAM guide to provenance research*. Washington DC: American Association of Museums.

For Peer Review Only