A Lightweight Authorization Mechanism for Spatially Enabled Health Data Services

Martin Tomko[+]
tomkom@unimelb.edu.au

Christopher Bayliss[+]
baylissc@unimelb.edu.au

Gerson Galang[+]
ggalang@unimelb.edu.au

Tristan Chadwick*
tristan.chadwick@health.wa.gov.au

James Cosford*
james.cosford@health.wa.gov.au

Richard Sinnott[+]
rsinnott@unimelb.edu.au

[+]The Australian Urban Research Infrastructure Network
The University of Melbourne
VIC, 3052 Australia

* CRC – SI Project Team
WA Department of Health
189 Royal St
Perth, WA, 6004 Australia

ABSTRACT

Ensuring that only authorized users have access to certain *sensitive* datasets is of paramount importance and this is especially so in the health sector. While the importance of the ability to access and utilize such data to better manage public health has been increasingly recognized, the process of defining and enforcing access management remains largely *ad hoc* with data provider specific solutions typically required. This is due to the heterogeneity of data and in situations where systems are already in place and are expected to remain so for a foreseeable future. In this context, we present a *lightweight* and *data provider-driven* software system for providing access to health records that include geospatial information. The proposed architecture is lightweight as it focuses on the re-use of existing data-access standards and services familiar to health data providers. The system is *data provider-driven* since it only allows access to and use of health data for research purposes through data provider initiated processes. The solution described has been designed in the context of the Australian Urban Research Infrastructure Network (AURIN) Project together with a (federated) data provider to extend the visibility and accessibility of their spatially enabled healthcare database.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Spatial databases and GIS*;* H.3.5 [**Information Storage and Retrieval**]: On-line Information Services.

General Terms

Design, Security

HEALTHGIS '13, November 05 - 08 2013, Orlando, FL, USA

http://dx.doi.org/10.1145/2535708.2535719

Keywords

eResearch, Federated Data, Urban Research, Authorisation, Ethics.

1. Introduction

Individual health records present a treasure trove of information that can provide important evidence about the state and change of the health of the population. At the same time, such records are amongst the most sensitive data held by any public agency or authority. The access to such data mandates that only authorized individuals should be allowed. Here authorized often implies that they have ethical clearance to do so and that the data providers/authorities have agreed that access is allowed through their own internal processes, which can include obtaining patient consent. Access and use by unauthorized users could lead to their leakage, abuse and a decrease in trust by the agencies involved [2]. This, in turn, could lead to the reduced ability of public health agencies to monitor and manage the health of a nation, with dire consequences. The potential for abuse of health records is further increased if the records are stored at individual (patient) level and are either directly or indirectly spatially referenceable (e.g., linked to an address).

Advanced systems with sophisticated authorization and authentication mechanisms enabling access to health records are so far rare, and usually focused on mission-critical applications in the management of individual health records by the administration or in limited-scale clinical trials [6]. At present there are no integrated national data access systems for health data across Australia or indeed internationally. Many of the challenges in supporting such ubiquitous systems are typified by the UK Connecting for Health National Program for IT and in its ultimate failure in delivery of the integrated national platform [7]. Health geography researchers have long been experiencing difficulties in gaining access to health outcome databases, with access often negotiated on an individual research basis. Upon being granted ethical clearance, researchers would typically get access to a specific data extraction often delivered through *ad hoc* out of band mechanisms. We argue that this approach is not only inconvenient (the researchers may have troubles handling the data due to storage, computational capacity and other technological constraints), but ultimately this approach also reduces the agency's control over such information (in cases of theft, inappropriate handling, or it's use past any particular research

clearance period). With the advancements in Internet-based software engineering and accumulated experiences of the processes of access to and use of health data in a range of scenarios [13], this modus operandi of health data access can be tackled.

In this paper, we present a lightweight approach to ethical clearance-based authorization developed in the context of the Australian Urban Research Infrastructure Network (AURIN). AURIN is tasked with development and delivery of a national eResearch infrastructure offering a data access, analysis and visualization platform for urban and built environment researchers. Key to AURIN is that the data itself remains *in situ* at the data providers and the platform provides *federated* access to these data sets. There are many national data providers offering data resources through the AURIN Platform including, amongst others, the Australian Bureau of Statistics. AURIN provides a single unifying Portal offering a "lab in the Web browser" enabling Australian researchers across a range of disciplines access to federated data sources and tools. For more information on the AURIN project see http://www.aurin.org.au.

In this paper we first discuss the need for securitized access to (spatially referenced) health records (Section 2), and provide background on authentication and authorization systems used in data access. We then outline the main functional requirements demanded when health records are to be exposed to an eResearch infrastructure that aims at a minimal impact on the data provider (Section 3). We then discuss the approach proposed and developed (Section 4), in this case extending the data access capabilities provided by an Open Geospatial Consortium (OGC) Web Feature Service (WFS) compliant data service. We discuss the pros and cons of this approach and conclude the paper with a discussion of future extensions envisaged.

2. BACKGROUND

The access to protected health data records must be controlled in such a way that only *authenticated* and *authorized* users have access to the protected information. Here authentication of users implies that their identity has been established – typically through a username and password challenge-response mechanism. Authorization of users implies that they are allocated privileges that are subsequently used to determine what they are allowed to do and this is subsequently enforced – in the case of AURIN this is often related to data access and use demands put in place by the data providers themselves. Key to this in the context of AURIN is that research end users should never be able to establish the identity of individual patients. Furthermore, the processes of data access itself and implicitly trusting technology to define and protect data access is often fraught. Many data providers are unwilling to provide direct (programmatic) access initiated by incoming connections from the Internet, irrespective of whether authorization technologies are in place. There are several reasons for this including their lack of familiarity with more advanced web-based authorization technologies. On the other hand, spatially referenced health records provide some of the most powerful instruments for the analysis of health outcomes, and are therefore highly sought after by academic researchers and policy analysts. Privacy concerns need to be address in a particularly stringent manner when analyzing individual level health outcomes where privacy protection needs to be strongest, such as in pervasive healthcare [4].

2.1 Privacy Control

The access to spatially-referenced health outcome data enables the correlation with other environmental and social factors co-occurring, and possibly contributing either positively or negatively to the health of a population. Access to such research can extend the ability of authorities to mitigate the adverse development of public health, but needs to be counter-balanced by measures to assure individuals privacy [1; 8]. However with authentication and advanced authorization systems coupled with targeted usage scenarios, i.e. where agreed data sets are to be accessed, the access to and aggregation of information about health outcomes can be realized and together allow to overcome many of the direct patient privacy challenges [11; 14]. Patient privacy itself is best tackled at the source of the data (e.g. the hospital or health authority). A typical hurdle that must be overcome is the possibility of further linkage of individual level data and subsequent risks of statistical identification of patients. This risk is significant where data from the *same patient* (or individual in the case of government data) is being linked across multiple organisations. Solutions such as Vanguard have been developed with this scenario in mind [6; 7]. Anonymization techniques and use of advanced statistical disclosure risk control solutions represent other approaches that have been put forward and adopted. However disclosure control [14] approaches typically do not transfer across organisational boundaries since agencies (and hospitals) have different policies in place and these policies can only be truly tested once access to data and risk control is offered. Such data linkage and risk control policy testing is often realised through secure data archives where researchers have to physically go to a secure environment to access and use sensitive data sets. The Office of National Statistics Virtual Microdata Laboratory in the UK (www.adls.ac.uk) and the Sax Institute in Australia (www.saxinstitute.org.au) are two examples of such facilities. The limitations of such solutions are that researchers are requested to physically going to such facilities to access and use such data. In the context of urban research it is often unit level data that is aggregated to a given geospatial coverage that researchers are primarily interested in. In this paper, we focus only on the first two aspects of data access, authentication and authorization.

2.2 Access Control

Security systems have to be simple for users. Authentication through AURIN is achieved through the Australian Access Federation, which provides decentralized (federated) authentication. In this model researchers at particular academic institutions across Australia are able to log in to the Portal through their own institutional identity provider [5]. Authorization relates to the finer-grained control of access, for instance based on individual's roles and their matching with the access rights to the datasets. A multitude of authorization technologies exist – the most common being Role-based Access Control (RBAC). In this model privileges (roles) are assigned to users that are associated with particular access and usage policies. A common approach is to have policies coupling *roles*, *targets* (e.g. particular databases or database tables) and *actions*, e.g. read, write etc. The implementation of RBAC systems is greatly simplified when pre-agreed access and usage scenarios are implemented (as opposed to ad hoc, dynamic queries for arbitrary data sets from given providers). A review of authorization technologies is given in [13]. Within AURIN the association of privileges and their usage with authentication information has to be seamless, i.e. researchers are assigned privileges and these are used as and when required to access distributed data resources without further

challenge/response demands. This transparency is commonly known as single sign-on.

2.3 Data Access through Federated Spatial e-Infrastructures

The need for adequately designed access control systems is increasing in the era of multi-institutional collaborations and data mash-ups. eResearch Infrastructures present one of the tools facilitating this trend, framed in the context of e-Science [3]. Spatial data linkage and analysis enables novel insights in particular when combining datasets across institutional and disciplinary boundaries. Health data can be interpreted in the context of environmental and social datasets, often curated by different agencies. Spatially-enabled eResearch profits from aspects of CyberGIS tools to achieve such data linkage and analysis [12; 15]. The AURIN project has developed one such eResearch infrastructure, providing urban researchers with access to federated data sources that can be further visualized, combined and analyzed in a lab-in-a-browser environment [9; 10]. Security and data control are paramount if the trust of data curators is to be maintained and datasets opened for access. A variety of security approaches has been explored in the context of AURIN, such that they suit the particular needs of individual data providers [5]

3. FUNCTIONAL CONSIDERATIONS FOR LIGHTWEIGHT SECURE ACCESS TO HEALTH DATA

In this section, we discuss the overall user interaction flow and the resulting technical realization supporting federated access to remote datasets, and relate these requirements to the specific conditions that need to be satisfied by such remote clients when accessing federated health data sources.

3.1 Requirements to Federated Access to Health Data Sources

The AURIN architecture presented has been designed with particular emphasis on reducing the impact on the existing infrastructures of data providers. Health data providers are exposing their data to the infrastructure with the following core restrictions:

1. **Authentication**: the authentication of the users is to be handled by the client, with a common identifier passed to the authorization system;
2. **Authorization**: the solution must enable the data provider to authorize a user's access to the data based on their current ethics clearance process with minimal impact on existing agency solutions;
3. **Isolation**: the access to the data cannot expose a (web) service directly to requests incoming from the Internet. The software deployed on the health data providers' infrastructure must be kept isolated in order to assure data security and avoid other costly measures (e.g., penetration tests). The active part of the communication (the request) must be based on an action initiated by the data custodian;
4. **Interoperability**: the solution should enable the data provider to use their current internal data dissemination service and/or an off the shelf, standardized data provision solution (such as an OGC WFS service in case of spatially referenced data) for ease of integration with existing data clients.

This set of requirements often poses a challenge to the realization of the data access.

3.2 Functional Requirements of a Federated eResearch Platform Client

The AURIN Portal is an eResearch platform accessing (as a client) various federated data sources. It has been designed to support the interaction flow from data discovery to knowledge creation to knowledge sharing. The user can search for available datasets based on harvested (and enriched) metadata records and specify queries requesting subsets of the federated datasets based on spatial, temporal and attribute filters. The data specification is then transformed into the particular type of request matching the data source from which the data are accessed [9; 10].

In the context of a generic, federated data access platform, access to securitized health datasets should require minimal changes of the platform's architecture and seamlessly add to other data clients deployed. The solution should satisfy the following characteristics:

1. **Authentication**: the authentication of the data client (through the federated data access platform) should not be bound to specific user's credentials. Rather many users should be able to access and use these data access services. For all purposes, the requests would be arriving from the same client.
2. **Authorisation**: the client must be able to lodge data requests to the data service by supplying identifying information about the requesting user that is commonly available to the client system, and to ensure that the user can be matched with an independently maintained authorization database at the data provider's end. This authorization information should be available to the client, but hosted within a secure repository associated with the federated data access platform.
3. **Long-runtime asynchronicity**: The client must be able to lodge a user's data request any time, and be able to receive the response anytime, thus they should not be impacted by delays at the data provider. This is in particular important in situations where the client is not able to test the status of the remote data service and/or assert whether a given user has been granted access.

4. TECHNICAL REALISATION

In this section, we describe the overall interaction workflow and summarize how the main elements of the realised solution help to satisfy the functional requirements mentioned.

4.1 High-Level Interaction Workflow

We briefly outline the sequence of interactions between the client and service systems (see Figure 1). Only the main sequence of interactions between the client and data service is described, with the interaction within the remaining components of the AURIN system not described. For more information about the complete sequence, we refer the interested reader to [10]. We start the description at the stage where the parameters required for the data request are received by the appropriate type of client (WFS client). All numbers below relate to Figure 1.

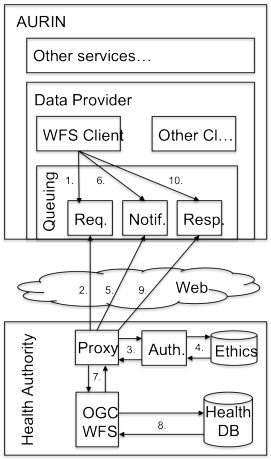

Figure 1 The interaction between the AURIN Infrastructure and the securitized health data provider.

1. The WFS client (or, more precisely, the Proxy-enabled WFS client, also known as *worker*) is the data client that receives the parameters for a given subset of data from a dataset. These parameters include filtering information, connection parameters, and, importantly, the user identification parameters used in authorization – in our case the user's email address. Here it is important to note that the user of the AURIN system accesses the environment through a single sign-on authentication system. In this model, user's identification parameters are provided by the their identity provider (e.g. their home university). As a single user may have a number of email addresses, this information is out of control of the AURIN system. We will discuss the ramifications of this constraint later. Once the parameters are received, the client formats the request message, and lodges it into a request queue.
2. The request is queued and identified by a unique job identifier. The request waits in the queue until the data provider's worker processes all the preceding messages (on a first in – first out basis) and picks the request. At this point, the authentication of the client system is assured – the data provider initiates the request with their own authentication parameters and perceives the data client as a single entity for all requests. The key point here is that the request query is initiated on the data provider side.
3. The request message is then parsed by the data provider's proxy and the user's email address sent to the authorization service.
4. The authorization service checks the user's email address against the ethics approvals database to determine whether the address is known. The (positive or negative) result of the authorization is communicated to the Proxy service.
5. The Proxy service lodges a notification in a notification queue, either informing the client that the request is being honored and is being processed, or if the authorization was unsuccessful, informing the client that the user's email address is not present in the data agency's authorization database and subsequently further providing information how to request access.
6. The notification queue is regularly checked by the data client regarding the status of requests that have been issued. The failure of a request is immediately propagated to the core system for user notification. The success of a request (processing / done status) are used by the data client (in this case a WFS client) to determine its future actions (see point 9).
7. After the user is successfully authorized and a positive notification is propagated to the client, the request is transformed (if needed) into a format acceptable by the data service (in this case, an OGC WFS service). This transformation is obfuscated from the client and only happens in situations when a bespoke API is queried and the data provider does not desire to disclose its details to the client.
8. The data are retrieved form the backend database. At this stage, additional securitization may happen, for instance performing checks whether the values reported satisfy privacy constraints (such as minimum counts of disease incidents per region and others).
9. The results of a successful request are formatted into the agreed response format and lodged into the response queue, from where they are retrieved by the client system and either directly passed further to the internals of the e-Infrastructure, or additional validation and formatting processes are applied to, for instance, transform the response into a standardized internal representation acceptable by the e-Infrastructure.

It is worth noting that the above interaction can also be run in a simplified setup in the case that advanced notification is not required. Then, the interaction through the notification queue is not necessary, and the response queue can hold both successful results (the data) and error responses if the request fails.

4.2 Publish/Subscribe Architecture and Message Queuing

The Message queuing component of the architecture allows decoupling of the internal realization of the data service from the client environment, and provides the data provider with full control over the access and load on its infrastructure. The queuing subsystem does not introduce substantial additional latency into the processing chain. The currently used system (the Open Source

system ActiveMQ, http://activemq.apache.org/) has been designed to support time-critical applications, such as high frequency trading, and is therefore suitable for use also in non-real time critical application scenarios. The data volume transferred in the case of spatially-enabled health data records is substantially higher than in other domains. Our tests show that this is not a problem for the current implementation, in particular in the case of aggregate-level data and if geometries are not requested (geometries are already available and ingested into the AURIN core e-Infratsructure). The main time lag is introduced by the authorization and data retrieval parts of the interaction, and these are identical to those required in direct WFS access.

Long-runtime asynchronicity is also assured by the queuing systems. Should the e-Infrastructure, data client (OGC WFS client in the case above) or the data provider itself experience intermittent outages, the messages will wait in the queue and can be delivered with longer delays. This is also convenient in situations where the ethics approval process is contingent not only on automatic check against a whitelist user database, but possibly also on manual approval by one or more data custodians.

4.3 Proxy-based Gateway - Authorization and Message Manipulation

The necessity to provide an active service that interacts with the request, notification and response queues provides the opportunity to add additional logic to the request and response validation, interpretation, and transformation. The interaction with the user registration database providing the authorization backend is one of the main functionalities attached to this proxy component. Another positive is the ability to decouple the data service's internal API from the API documented to the client. Indeed, in a number of realisations, we lodge request formatted as JSON messages based on agreed key-value pairs containing request parameters, and let the final formatting (and likely enrichment) of the request to fit the legacy format to the data provider's Proxy. Similarly, the responses from the data service may be further manipulated before lodged in the response queue.

4.4 Common Authorization Identifier

The use of the normal user email address as the common as the identification parameter may seem inappropriate, especially in the light of more advanced authorization schemes, such as those discussed in [5]. The difference between these architectures and the one proposed is the low impact on the data provider's (and, ultimately, the client e-Infrastructure's) extant technical setup. The deployment of a more sophisticated authorization system based on encrypted shared user tokens would require substantial re-engineering of the data provider's software solution and would also require the administration of an additional user registration system. Furthermore, the eResearch platform would not be able to use the standard single sign-on infrastructure provided to the academic community that is disconnected form that of the health data providers, and without a possibility to maintain a shared control over the user tokens. The proposed solution further provides transparency to the user about the reasons that their request is not granted – the used email address is a human readable string that is propagated from the user's Identity Provider (IdP). The user can take all action necessary to have additional email addresses added to the ethics approvals database of the data provider if needed.

5. DISCUSSION

The proposed approach, as implemented and tested in AURIN, satisfies the functional requirement for a lightweight, minimally intrusive securitized access solution to an existing spatially enabled dataset of health outcomes, in particular in the context of records with aggregate-level attributes relating to relatively large spatial regions (health districts and other statistical regions). It provides researchers with access to the most current records generated based on the most recent records held by the health agency. Furthermore, the data provider maintains continuous control over their datasets, for instance maintaining a real ability to expire an approval of access. This has been traditionally difficult as authorized researchers were usually handed a DVD with a snapshot of the data.

The approach documented has only been tested on aggregate-level datasets. These relate to a relatively limited number of regions within a country. It is yet to be seen how a queuing-based approach will cope with potentially much larger datasets coded at individual record levels. It is, however, not a pressing concern as getting access to such datasets is highly problematic and likely out of scope of our current project.

The isolation of the data service behind a proxy service poses additional challenges, such as the monitoring of the heartbeat of the data provider's system (availability and response time monitoring). We are currently using a modified set of queues to assure this functionality, but we are end without control should the data provider's queue listeners become unavailable. On the other hand, this provides the data providers with assurance that they are fully in control of access to their data services.

6. CONCLUSION AND FUTURE WORK

We have presented an approach to securitization of existent healthcare databases through a lightweight system that utilizes (spatial) data infrastructures for accessing federated and sensitive health-related datasets. The approach has been realized in AURIN with two state health data providers in Australia (VicHealth in Victoria and Department of Health in Australia). Large amounts of initially aggregate level datasets are, for the first time, exposed to a large range of urban researchers.

AURIN aims to provide a flexible platform for urban research beyond just public health. On the contrary, the aim is to provide a single point of truth for a large amount of urban-related datasets that can be cross-referenced, combined, and analyzed in context. It is therefore important to devise approaches that minimize the impact of any single data provider on the entire e-Infrastructure, while providing data providers with integration options that will not be perceived as burdens that stand in the way of their day-to-day operations. We believe that the approach discussed is one such solution. The currently implemented system allows access to the securitized datasets and enables their visual exploration through dynamically linked mapping and charting capabilities, as well as hypothesis testing through a growing range of hypothesis testing statistical tools, including a rich spatial statistics toolset. We are currently exploring how this approach can be further expanded to provide the ability to analyze user-provided datasets in the context of highly securitized, unique-records through the ability to send the processing logic and part of the input datasets to the data provider.

7. ACKNOWLEDGMENTS

AURIN is an Australian Government project conducted as part of the Super Science initiative and financed by the Education Investment Fund.

8. REFERENCES

[1] Gao, S., Mioc, D., Yi, X., Anton, F., Oldfield, E., and Coleman, D.J., 2009. Towards Web-based representation and processing of health information. *International Journal of Health Geographics 8*, 3. DOI= http://dx.doi.org/10.1186/1476-072X-8-3.

[2] Gutmann, M.P. and Stern, P.C., 2007. *Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data*. The National Academies Press.

[3] Hey, T., Tansley, S., and Tolle, K., 2009. The Fourth Paradigm: Data-Intensive Scientific Discovery Microsoft Research, Redmond, WA, 284.

[4] Moncrieff, S., Venkatesh, S., and West, G., 2009. A Framework for the Design of Privacy Preserving Pervasive Healthcare. In *IEEE ICME 2009* IEEE Xplore.

[5] Sinnott, R., Bayliss, C., Galang, G.G., Damien, M., and Tomko, M., 2012. Security Attribute Aggregation Models for e-Research Collaborations. In *Trust, Security and Privacy in Computing and Communications (TrustCom 2012*, M. Geyong and M. Felix Gomez Eds. IEEE, Liverpool, UK, 342 - 349.

[6] Sinnott, R.O., Ajayi, O., and Stell, A.J., 2009. Data Privacy by Design: Digital Infrastructures for Clinical Collaborations. In *International Conference on Security and Privacy* FI, USA, Orlando.

[7] Sinnott, R.O., Stell, A.J., and Jiang, J., 2011. Classifying Architectural Data Sharing Models for e-Health Collaborations. In *HealthGrid 2011*, Bristol, UK.

[8] Theseira, M., 2002. Using Internet GIS technology for sharing health and health related data for the West Midlands Region. *Health & Place 8*, 1, 37-46. DOI= http://dx.doi.org/10.1016/S1353-8292(01)00034-X.

[9] Tomko, M., Bayliss, C., Widjaja, I., Greenwood, P., Galang, G.G., Koetsier, G., Sarwar, M., Nino-Ruiz, M., Mannix, D., Morandini, L., Voorsluys, W., Pettit, C., Stimson, R., and Sinnott, R., 2012. The Design of a Flexible Web-based Analytical Platform for Urban Research. In *ACM SIGSPATIAL '12 Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, I.F. Cruz, C. Knoblock, P. Kroger, E. Tanin and P. Widmayer Eds. ACM Redondo Beach, California, USA, 369-375. DOI= http://dx.doi.org/10.1145/2424321.2424368.

[10] Tomko, M., Galang, G.G., Bayliss, C., Koetsier, J., Greenwood, P., Voorsluys, W., Mannix, D., Sarwar, M., Widjaja, I., Pettit, C., and Sinnott, R., accepted. Designing Adaptable Spatial Cyberinfrastructures - the example of a Loosely-Coupled Internal System Architecture for Urban eResearch In *CyberGIS: Fostering a New Wave of Geospatial Discovery and Innovation*, S. Wang and M. Goodchild Eds. Springer Verlag.

[11] Vanwey, L.K., Rindfuss, R.R., Gutmann, M.P., Entwisle, B., and Balk, D.L., 2005. Confidentiality and spatially explicit data: Concerns and challenges. *Proceedings of the National Academy of Sciences of the United States of America 102*, 43, 15337-15342.

[12] Wang, S., 2010. A CyberGIS Framework for the Synthesis of Cyberinfrastructure, GIS, and Spatial Analysis. *Annals of the Association of American Geographers 100*, 3, 535-557.

[13] Wei, J., Arshad, J., and Sinnott, R.O., 2011. A Review of Grid Authentication and Authorization Technologies and Support for Federated Access Control. *Journal ACM Computing Surveys 43*, 2. DOI= http://dx.doi.org/10.1145/1883612.1883619.

[14] Willenborg, L. and De Wall, T., 1996. *Statistical Disclosure Control in Practice*. Springer, Berlin.

[15] Yang, C., Raskin, R., Goodchild, M., and Gahegan, M., 2010. Geospatial Cyberinfrastructure: Past, present and future. *Compters, Environment and Urban Systems 34*, 4, 264-277.