Secure Data Sharing Polices and Architecture Preserving Privacy

Sanaa Sarahneh
Palestine Polytechnic University
sanaa.s@ppu.edu

Radwan Tahboub
Palestine Polytechnic University
radwant@ppu.edu

***Abstract*—Electronic data interchange can be classified as one of the important areas of information technology, where the need for data sharing increasingly required in almost every field. Data sharing concept can be defined as the process of interchanging, analyzing, retrieving and integrating data among multiple data sources in a controlled access manner. The use of information technology in different areas began to increase since 1980s; the exchange and sharing different types of information was required at that time. Although data sharing facilitates the way that data can be exchanged, security concerns arise as a challenge for conducting data sharing, many polices include confidentiality and privacy must be taken into consideration. This study will provide a literature review of security policies, focusing on privacy models to facilitate data sharing among different organizations in different areas.As a result for the study there are diffrent data sharing model that applies diffrent polices to preserve privacy such as Semantic Privacy-Preserving Model, Capability-based Access Control Model, and OneSwarm data sharing Model.**

***Index Terms*—Data Sharing, Privacy, Security, Access Control, Management, Policies.**

I. Introduction

Nowadays, most organizations expanded their work in the form of extranet to facilitate exchanging data among each other.Electronic data interchange can be classified as one of the important areas of information technology, where the need for data sharing increasingly required in almost every field. Data sharing can be defined as the process of interchanging, analyzing, retrieving and integrating data among multiple data sources in a controlled access manner. The use of information technology in different areas began to increase since 1980s; the exchange and sharing different types of information was required at that time. Although data sharing facilitates the way that data can be exchanged, security concerns arise as a challenge for conducting data sharing.

The remainder of this paper is organized as follows: the next section provides background in ata sharing and security in data sharing. Section 3 explains three models that preserving privacy in data sharing. Next, section 4 compares between the models in terms of their advantages and disadvantages. Finally, we conclude in Section 5.

II. Background

This section provides data sharing concepts, the need for data sharing, data sharing management, and the security for data sharing.

A. *Data Sharing Concept*

Data sharing concept emerges to introduce a new era of cloud computing processes, e-commerce, e-government, e-operations, e-everything. This term was coined since 1970s as Bakis et al., (2007) indicate. Bakis et al., (2007) add from the early 1980s, the use of IT in the construction industry and broader engineering sector began to increase and find application in many different areas, the exchange of many different types of information was required at that time. Sarathy and Muralidhar, (2004) also describe data sharing as a fundamental enabler of coordination among supply chain partners. Therefore, data sharing can be defined as the process of interchanging, analyzing, retrieving and integrating data among multiple data sources in a controlled access manner, also Greif and Sarin (1987) define data sharing as a fundamental to computer-supported cooperative work; people share information through explicit communication channels and through their coordinated use of shared database.

B. *The Need for Data Sharing*

Sarathy and Muralidhar (2004) find out that data sharing is an important feature for modern organizations due to the increase in the use of communication networks, changes in architectures of enterprise information systems, as well as the increasing availability of data in computerized form, and perhaps the biggest impact on data sharing can be attributed to the widespread use of the Internet and Internet-related technologies for e-government, e-commerce, scientific research and healthcare. They add, e-government involves sharing data for transactions with citizens, other agencies and outside vendors and businesses. Sarathy and Muralidhar (2004) add, in e-commerce, data can be shared for transactions, operations, and analysis. Conducting business transactions is a basic reason for sharing data in e-commerce it is mainly used in Electronic Data Interchange (EDI), business to-business marketplaces, as well as consumer purchases over the web. They empathize that the focus of data sharing for operational purposes leads to the optimization of business processes over the entire chain to benefit all participants in the chain. Information that shared among

supply chain partners may include inventory sales, demand forecasts, order status, and production schedules. Analysis, business intelligence, and decision-support represents the third purpose for data sharing in e-commerce, information available for analysis is increased through the sharing of data, they provides an example of banks data sharing with affiliates and telemarketers, another example about retailers who allow suppliers to access their inventory data for analysis purposes. Where Mannai and Bugrara, (1993) indicate that it is highly desirable to share data among the members of the medical community; because data is very valuable, hard to produce, and in some cases irreproducible resource. Data sharing reduces the cost of reproducing redundant data collections as much as minimizing the efforts paid in performing this.

C. *Data Sharing Management*

Since data sharing coined, emerging data from heterogeneous sources into a single common to make data compatible with each other becomes critical issue as Harris et al., (2007) indicate.Data integration has been attempted for about 20 years, Hu and Yang, (2011) define data integration as the problem of combining the data from autonomous and heterogeneous sources, and providing users with a unified view of these data through. Varlamis and Vazirgiannis, (2001) add that many organizations and enterprises establish distributed working environment, where different users need to exchange information based on a common model, XML (eXtensible Markup Language) is used to facilitate this information exchange. The extensibility of XML allows the creation of generic models that integrate data from different sources and XML is becoming the standard format for data exchange among distributed applications components. The use of XML for information interchange among different enterprises and organizations evokes the need for common schema that the information must follow.

D. *Secure Data Sharing*

Although data sharing facilitates the way that data can be exchanged, security concerns arises a challenge for conducting data sharing, confidentiality and privacy must be taken into consideration, this means, a controlled access is required to authorize authenticated users or roles to access data. Each data source represents a database, each database may use an application -for example- to access another database, this application is assigned specific permissions to access specific view of a specific database, permissions that identifies what kind of access must be granted to this application, (e.g. to read, or write, or even to have full access), for this purpose, a database of databases is needed to allow the sharing of data among the different databases as Mannai and Bugrara, (1993) indicate. Clifton et al., (2004) say that this increase the need for data sharing management and data integration, on another hand data sharing and integration are prevented from being widespread because of privacy concerns, for example in e-commerce areas companies need to exchange information to boost productivity, but are prevented by fear from competitors, also sharing data in healthcare areas improve scientific research and enables early detection of disease, but without preserving privacy it is costly and difficult to make healthcare information globally expand. Isdal et al., (2010) define privacy as the process to protect information from unauthorized access. Harris et al. ,(2007) say that cyber crime as well as threats to national security is costing organizations billions of dollars each year, it is equally certain that unrestricted data sharing will reduce the privacy and/or confidentiality of individuals, Harris et al., (2007) add the challenge is to enforce appropriate administration and security policies that facilitate data sharing as needed .These policies include policies for confidentiality, privacy, and trust. During normal operations, it is important to maintain confidentiality and privacy. In addition, trust policies ensure that data is shared between trusted individuals. The standards efforts in this area include Role-based access control (RBAC) as well as Platform for Privacy Preferences (P3P) Choi et. al. (2013),also add that Public Key Infrastructure (PKI): preventing illegal modification, edits, or transfers of sensitive data to a third parties for unintended purposes. . Different models have been introduced to apply privacy in data sharing and data integration, each may be the same or different structure of other, the next section provide a literature review for preserving privacy models in different areas.

III. Preserving Privacy Models in Data Sharing

This section provides a literature review for preserving privacy models for data sharing in different areas, and their Strengths and weaknesses.

A. *Semantic Privacy-Preserving Model*

A semantic privacy-preserving Hu and Yang, (2011) model provides authorized view-based query answering over a widespread multiple servers for data sharing and integration. For that reason model consider a large number of servers. Therefore a unified global data sharing and protection service can be achieved at the virtual platform (VP).

1) The combined semantics-enabled privacy protection policies are used to empower the data integration and access control services at the (VP). Privacy protection policies represent a long-term promise made by an enterprise to its users and is determined by business practice and legal concerns, which is expressed as combination ontology and rule:
   - A privacy protection policy is a type of formal policy (FP) used for specifying a data usage constraint from a data owner. FP is a declarative expression corresponding to a human legal norm that can be executed in a computer system without causing any semantic ambiguity.
   - An FP is created from a policy language (PL), and this PL is shown as a combination of ontology language and rule language.
   - A formal protection policy (FPP) is an FP that aims at representing and enforcing resource protection

principles, where the structure of resources is modeled as ontology's O but the resources protection is shown as rules R.(It is combination of ontology's and rules O+R)
- Semantic Web Rule Language (SWRL) Tab development tools and Semantic Query-Enhanced Web Rule Language (SQWRL), Web Ontology Language (OWL-DL) query language to model and enforce semantic privacy protection policies.

2) Three approaches have been proposed to model a set of source descriptions that specify the semantic mapping between the source schema and the global schema:
   - Global-as-view (GAV) requires that the each concept in the global schema is expressed in terms of query over the data sources.
   - Local-as-view (LAV) requires the global schema to be specified independently from the sources, and the source descriptions between the stable global schemas.
   - Global-local-as-view (GLAV), a source description that combines the expressive power of both GAV and LAV, allowing flexible schema definitions independent of the particular details of the data sources.
3) This model is proposed with three layers, where the bottom layer provides data sources from the relational databases .The middle layer provides a semantics- enabled local schema for each independent service domain. The top layer is served at the VP, which provides a unified global view of privacy-preserving data sharing and integration services.
4) The ontology mapping and merging algorithm with a local-as-view (LAV) source description that creates a global ontology schema (mediated), which is a reconciled view of the information that provides query services to end users ,at the VP by integrating multiple local ontology schemas for data sharing. Model merged global ontology schema that mentioned above in the middle layer.
5) Using description logic (DL) to model the local and global schemas is to empower the ontology's abstract Concept representation and reasoning capabilities.
6) A query is defined as an SQWRL data log rule in the SWRL-based policy to access to a global ontology, and each SQWRL data service query for a global ontology at the VP is mapped to multiple queries as SQWRL data log rules for each local schema.
7) The challenge of designing a semantic privacy protection model is to ensure soundness and a completeness of data sharing and protection in multiple servers:

- For the soundness criterion, this model does not allow unintended data being released to the data users through the global policy schema (GPS) at the VP.
- As for the completeness criterion, the model does not miss any eligible shared data when a user asks for a data request service at the VP. Therefore, shareable data obtained the VP should equal data obtained directly from each server.

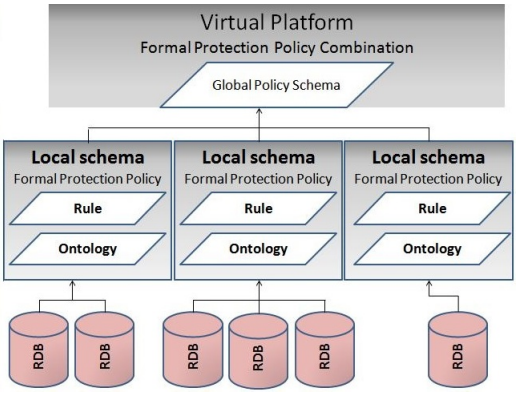

Figure 1. **A semantic privacy protection model. Source: Hu and Yang, (2011).**

Figure (1) is proposed with three layers, where the bottom layer provides data sources from the relational databases (RDB), the middle layer provides a semantics- enabled local schema for each independent service domain. The top layer is served at the VP, which provides a unified global view of privacy-preserving data sharing and integration services. In the top layer at the VP, we have a global policy schema (GPS), including a global ontology schema (GS) aligned and merged from several local schemas (LS), e.g. TBox and a set of rule integration at the middle layer. The VP provides conceptual data access and protection services that give users a unified conceptual-global view" with access control power for each data request. Ontology-based data sources are external, independent, and heterogeneous, and each local ontology was combined with logic program (LP)-based rules for each server in the middle layer. Mapping language (ML), which semantically links a GS and integrated rule, set in the top layer to each server's ontology LS and privacy protection rules in the middle layer. Ontology and the dynamic data sources are established by defining each concept in the data sources as a view over the global schema.

- **Semantic Privacy-Preserving Strengths:**

Hu and Yang, (2011) list some features in semantics privacy preserving model, First ,each server shares its collected data with other servers but without breaking the original data usage commitment to its clients ,therefore a unified global data sharing and protection service can be achieved at the virtual platform (VP). Second ,the model solve the soundness and completeness of query rewriting problem using a perfect ontology merging and a perfect rule integration from the local formal protection policies, For the soundness criterion, we do not allow unintended data being released to the data users. As for the completeness criterion, we do not miss any eligible shared data when a user asks for a data request service at the VP, Third, the model develop a privacy management framework and a formal semantics language to empower agents to enforce privacy protection policies. These formal

policy using ontology for privacy protection concept descriptions and rule for data query and access control services. Ontology-based data integration in DL is to provide a uniform access mechanism to a set of heterogeneous relational database sources, freeing the user from having the knowledge about where the data are, what they are stored, and how they can be accessed.

- **Semantic Privacy-Preserving Weaknesses:**

In spite of these features, this model still have a weaknesses, it face a background policy inconsistency problem when default policy assumptions vary between different servers (one server uses open policy assumption, where no explicit option-out for data usage mean option-in, but the other server uses closed policy assumption, where no explicit option-in for data usage means option-out) and to avoid this kind of policy inconsistency by requesting all sites to use a uniform policy assumption, and to collect option-in data usage choices from users whenever multiple policies are integrated. As a conclusion Semantic Privacy-Preserving model provide secure sharing through authorized views, each organization enables data sharing and data integration without affecting its clients, but the model have inconsistency problems.

B. Capability-based Access Control Model

Geambasu et al., (2007) use a model for data sharing called Capability-based Access Control, each capability consists of a Name, which identifies a single object in the internet, and group of access rights for that object. In this model, the system sits between applications and the underlying file system. It presents applications a view-based interface to the file system. It executes queries over the local file system and communicates with other peers to evaluate distributed queries. The model is depicted through the following steps:

1) The system registers each new view and capability in a local catalog, this capability has three parts (Figure 2):
   - A 128-bit global view Identification ID: this ID created by concatenating a hash of the local node's Media Access Control address (MAC address) with a locally unique-for-all-time view ID, this view ID uniquely identifies an individual view in the Internet.
   - A 128-bit random password: associated with each capability a 128-bit random password that ensures the capability's authenticity.
   - A 32-bit IP hint field: that contains the IP address of the node that likely contains or can locate the object addressed by the capability in the Peer to Peer Network (P2P), in general, they expect that objects will not move in their network, and the IP hint will be the address of the node that created the capability and still holds its definition. If the hint fails, then it must fall back on a conventional distributed hash-table scheme for location.
2) The per-node catalog table generated by the system holds view and capability information. It contains two tables ViewTable and CapTable(Figure 3). The ViewTable entry contains the global view ID, the view definition, and other attributes (such as the human-readable view name). For each view created on a node, there is one entry in a local view table (ViewTable). The CapTable entry stores the global view ID of the named view, the password, and the access rights. A node's capability table (CapTable) contains one entry for each capability minted to a locally known view.

128 bits	128 bits	32 bits
Global view ID	Password	IP hint

Figure 2. **Capability for a view. Source: Geambasu et al., (2007).**

Node-local view table (View Table)

Global view ID	View definition	Other attributes
...	...	...

Node-local capability table (CapTable)

Global view ID	Password	Rights
...	...	...

Figure 3. **Capability and view Catalog tables. Source: Geambasu et al., (2007).**

3) Users grant each other access to their data simply by exchanging capabilities to their views, much like users share access to private web pages by exchanging URLs.
4) When the system receives a capability, it uses the IP hint to determine whether the capability is for a local view. If the capability is local, the system checks whether the ¡global view ID, password¿ pair in the capability matches a ¡global view ID, password¿ pair in CapTable. If so, the capability is valid, and the system then examines the access rights in CapTable to see if the requested operation is permitted. If the capability is not found in CapTable or the operation is not permitted, the request fails. If the capability is for a remote view, the system forwards the request to the appropriate node in the peer-to-peer network, which then performs the validation itself.
5) To revoke a capability, the system simply removes an entry from the CapTable. Once a capability is revoked, all queries issued on that capability will fail.

- **Capability-based Access Control Strengths:**

Geambasu et al., (2007) add that Capability Based Access control model is a flexible protection mechanism for controlling access to shared views. Capabilities also enable rewriting and optimization of distributed queries, leading to good query execution performance. They also add, because capability is independent of the person using it, the systems access control scheme requires no user identities. Thus, sharing in a capability-based model requires no user accounts, no

user authentication, and no centralized protection structure. Capabilities facilitate data sharing because it can easily pass from user to user as a way to grant access.

- **Capability-based Access Control weaknesses:**

After revoking a capability, all queries issued on that capability will fail. But if a user with a capability has made a local copy of the shared data, revoking the capability cannot prevent him from distributing that copy. However, it prevents the holder from executing a query and seeing new or modified files that would result from that query. As a conclusion, the capability-based access control model provides flexible protection mechanism for controlling access to shared views, reuse of queries, it is independent of the user and decentralized.

C. Privacy-Preserving P2P Data Sharing with OneSwarm

OneSwarm (Isdal, et al., 2010) is a new P2P (Peer to Peer) design for data sharing that overcomes the lack of privacy in P2P data sharing applications such as BitTorrent- BitTorrent is an application that provides good performance but poor privacy- and to overcome poor performance in anonymizing overlays such as Tor. OneSwarm made a tradeoff between privacy and performance; it provides better privacy than BitTorrent and better performance than Tor. OneSwarm builds trusted links through social network peers, instead of relying only on a directory service such as a "Tracker" that gives information to the peers about the file. OneSwarm users are free to control the tradeoff between performance and privacy by managing the level of trust.

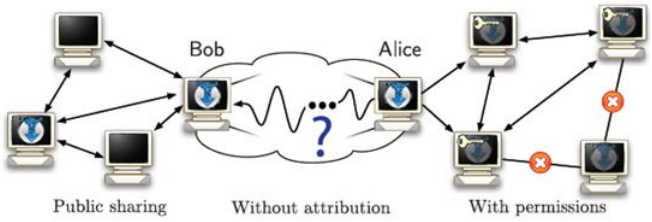

Figure 4. **Cases for data sharing by OneSwarm. Source: Isdal et al., (2010).**

There are three cases for OneSwarm described by Isdal, et al., (2010) and shown in Figure 4, the first one is public distributed data in this case the data is not private, and direct transfers between a large set of replicas yield. The second is sharing data with permissions limits access. The last one, data shared without attribution is accessible by everyone. In public distribution anyone in the network can download file free, all data is not private, and serves as fully backwards compatible BitTorrent client. With permission case only users with permission can download files, uses persistent identities to define per file permission, this case allows all acceptable users to recognize one another. While without attribution case is depend on obscuring attribution of source and destination, it uses privacy preserving keyword search, data is relayed through unknown number of intermediaries, and it is for sensitive material. The topology for OneSwarm the users define the links by exchanging public keys, this identifies each user and creates direct encrypted P2P connections, also OneSwarm uses social graph and community server for key distribution, Distributed Hash Table (DHT) serves as name resolution service, each client maintains encrypted entities advertising their IP address and port to authorized users, the topology is used for each transfer. In each transfer each OneSwarm client restricts direct communication to a small number of persistent contacts and locates different data sources using object lookup through overlay, this topology is used to enhance privacy, while to enhance performance in OneSwarm protocol, multiple paths to each data source are used. Linking peers with trust relationships is explained by Isdal, et al., (2010) it uses 1024 bit RSA (Rivest-Shamir-Adleman cryptosystem) public/private key pair which is generated in installation phase, public key serves as its identity among friends, manual key sharing between two users; the automatic key sharing discovers and exchange keys over local area network or by email invitation to friends. Managing untrusted peers by private community server and public community server, the private is to maintain a list of registered users and to provide authorized subscribes with a current set of public keys, the public is to allow new users to easily obtain a set of untrusted peers. Identity in OneSwarm protocol are managed by the DHT which contain of hashed IP and port, entries for a client encrypted with the public key, each entry is indexed by 20 byte randomly generated. Naming and locating data in OneSwarm used Secure Sockets (SSLv3) for connection as as Isdal, et al., (2010) say, file list messages is exchanged on first connection then compressed XML attributes which contain name, size and other meta data for particular peer. Shared files are named using 160 bit SHA-1 hash, for public data user obtains hashes from email, websites and keywords search, while for private data user must obtain both hash and key used for decryption of data. The risk in OneSwarm model as Isdal, et al., (2010) describe, the aattacker can join with limited number of nodes, also can check the traffic flow to/from, also may sniffing, modify or injected data. Limiting hacker to snoop in from by not assigning peer dynamically, also defining trusted and untrusted links to keep the information private, end to end path between users changes rapidly helps to prevent hacking using historical data. Isdal, et al., (2010) adds, preventing timing attack by search queries and responses are forwarded after adding a random delay to inhibit calculation of round trip time (RTT) to infer proximity, preventing correlation attack by having limited view of the overlay and cannot control path setup beyond directly connected neighbors, attackers could use this to correlate performance with ongoing transfers, finally preventing collusion attack by search queries and responses are forwarded probabilistically, making it very hard for directly connected colluding peers to infer source of data or monitor habits.

- **OneSwarm Data Sharing Model Strengths:**

OneSwarm provides flexibility for the user to manage the level of privacy for file sharing, incorporation of social network for building P2P file sharing network, and reduce cost of privacy.

- **OneSwarm Data Sharing Model Weaknesses:**

There are Delayed response to queries from untrusted peers.

IV. Comparison between the Models

The following table provides a comparison between the privacy preserving models in terms of their advantages and disadvantages:

Based on the comparison between the three models, the Capability-based Access Control model has disadvantages, mainly: there is no fixed method for translation, difficulties in integrity control, cannot prevent the user from keeping and distributing the shared data, and decentralized control.These disadvantages make the implementation of the model hard, concerning semantic privacy preserving model overcomes the previous disadvantages, and provides data integration, secure sharing through authorized view, in addition, each organization enable data sharing without affecting its clients, while OneSwarm data sharing model provides flexibility for the user to manage the level of privacy for file sharing, and reduces cost of privacy but it has delay in response.

V. Conclusion

Data sharing concept can be defined as the process of interchanging, analyzing, retrieving and integrating data among multiple data sources in a controlled access manner. Although data sharing facilitates the way that data can be exchanged, security concerns arises a challenge for conducting data sharing, many polices include confidentiality and privacy must be taken into consideration. In this study we provide a literature review of security policies, focusing on privacy models that facilitate data sharing among different organizations in different areas. As a result for the study there are diffrent data sharing model that applies diffrent polices to preserve privacy, and semantic privacy preserving model overcomes many disadvantages of others models, and provide data integration, secure sharing through authorized view, in addition, each organization enable data sharing without affecting its clients.

Table 1: Models Comparison

Model Name	Advantages	Disadvantages
Semantic Privacy-Preserving model	1. Each organization enables data sharing without affecting its clients. 2.Data integration. 3. Provide secure sharing through authorized views.	1.Inconsistency problems.
Capability-based Access Control model	1. Provide flexible protection mechanism for controlling access to shared views. 2. Reuse of queries. 3. Independent of the user.	1. Cannot prevent the user from keeping and distributing the shared data. 2. Decentralized control.
OneSwarm Model	1- Efficient, robust. 2- Users flexible.	There are delayed response to queries from untrusted peers.

References

[1] Bakis, N., Aouad, G., Kagioglou, M., (2007), "Towards distributed product data sharing environments Progress so far and future challenges", *Elsevier-Automation in Construction*, 16, (5): 586-595.

[2] Clifton, C., Doan, A., Elmagarmid, A., (2004), "Privacy Preserving Data Integration and Sharing", *ACM- Research issues in data mining and knowledge discovery*: 19-26.

[3] Geambasu, R., Balazinska, M., Gribble, S., Levy, H., (2007), "HomeViews: Peer-to-Peer Middleware for Personal Data Sharing Applications", *ACM*: 235-246.

[4] Greif, I., Sarin, S., (1987), "Data Sharing in Group Work", *ACM*, 5, (2): 187-211.

[5] Harris, D., Khan, L., Paul, R., Thuraisingham, B., (2007), "Standards for secure data sharing across organizations", *ACM-Computer Standards and Interfaces*, 29,(1): 86-96.

[6] Hu, Y., Yang, J., (2011), "A Semantic Privacy-Preserving Model for Data Sharing and Integration", *ACM-Web Intelligence, Mining and Semantics*, (9): 1-12.

[7] Isdal, T., Piatek, M., Krishnamurthy, A., Anderson, T., (2010), "Privacy-Preserving P2P Data Sharing with OneSwarm", *ACM SIGCOMM Computer Communication Review*, 40, (4): 111-122.

[8] Mannai, D., Bugrara, K., (1993), "Enhancing Inter-Operability and Data Sharing In Medical Information Systems", *ACM*, 22, (2): 495-498

[9] Sarathy, R., Muralidhar, K., (2004), "Secure and useful data sharing", *Elsevier*, 42, (1): 204– 220

[10] Varlamis, I., Vazirgiannis, M., (2001), "Bridging XML-Schema and relational databases A system for generating and manipulating relational databases using valid XML documents", *ACM*,: 105 - 114 .

[11] Choi, J., Chun, S., Kim, D., Keromytis, A., (2013), "SecureGov: Secure Data Sharing for Government Services", *The Proceedings of the 14th Annual International Conference on Digital Government Research*.