Trustworthiness of Linked Data Using PKI

Enayat Rajabi
Department of Technical and Engineering, Saveh Branch, Islamic Azad University, Saveh, Iran
erajabi@iau-saveh.ac.ir

Mohsen Kahani
Web Technology Lab, Ferdowsi University of Mashhad Mashhad, Iran
kahani@um.ac.ir

Miguel-Angel Sicilia
Information Engineering Research Unit, Computer Science Department, University of Alcalá, Alcalá de Henares (Madrid), Spain
msicilia@uah.es

ABSTRACT
The goal of this investigation is to enable users of Linked Data to assess the trustworthiness of datasets in order to decide if a piece of data fits their needs. The use of PKI[1] principles is proposed for trust management over the Web of Linked Data. Since datasets should exchange data in a trusted way, it is necessary for each dataset to have a digital certificate issued by a trustworthy third-party. This way, a trust relationship can be created directly between two organizations for data exchanging. Furthermore users can assess trustworthiness of datasets by identifying validity and quality of data via their certificates.

Keywords
Linked Data, Trustworthiness, PKI, Trust, Digital certificate

1. INTRODUCTION
The Web of Data is now rapidly being populated with thousands of datasets [1]. One of the goals of Linked Data is to provide useful information in RDF upon URI resolution [2]. Users can publish various open datasets on the Web and semantically relate data items from different data sources by using RDF links. When data is published from various resources, users may be facing the threat of receiving false and/or invalid or unreliable data on the Web or suffer from the problem that some of the URIs are not dereferenceable[2] . On the other hand, users always need to trust the information they search for through semantic search engines on the Web of Data. Further, trustworthy data has become much more critical when we face government data or statistical reports. While trustworthiness is usually posed in the context of active entities such as people or agents, there are few works about trustworthiness as a data quality criterion [3]. Some computer systems that use the trustworthiness and validity of data for filtering, usually apply a very simple assessment approach such as determining a trust score for the resources in which the data is related using one of the methods that exists for active entities even though sometimes we cannot map these kinds of methods to the Web of Data. However, the assessment of the trustworthiness of Linked Data can be a challenge as it used to be in traditional web. Here, we consider the trustworthiness of a resource exposed as Linked Open Data (LOD) using methods widely used for the Web.

In section 2, we consider the importance of Linked Data trustworthiness. Some different aspects of PKI and its influence on Web security will be discussed in section 3. In section 4, we will propose to use PKI principles for trustworthy assessment of datasets.

2. THE IMPORTANCE OF TRUSTWORTHINESS OF LINKED DATA

The Web of Data has been expanded and developed rapidly [4]. The rapid growth of the web of data and its worldwide adoption are fostered by the openness of the Web. This entails that datasets may derive from different resources. Since the Web of Data comes from diverse data sources with varying quality and different scope and assumptions, trustworthiness of data should always be well thought-out. Furthermore, any user can publish any kind of data without restrictions or controls, therefore poor quality data might be disseminated quickly through the interlinked data cloud and in the course of time we may have problems as fake datasets, deceptive data and invalid Wikipedia-extracted information. Many datasets and links may exist that are out of date and invalid. Users, as data consumers, need to have access to the updated datasets from their original sources. The credibility and validity of datasets should provide users of means to assess the trustworthiness of the websites, datasets, links, etc. As we mentioned before, there are also special data sources that require additional consideration. Government data, as an important resource, comes from different sources that each may have different quality, scope and domain. Another instance may be statistical and financial websites; it is obvious that when financial issues are discussed on the Web of Data, trust and credibility are of great importance because most of the resources are referred to as reference data.
There are several questions we should ask in this respect, e.g. who has created the content of the resource? Was the content ever manipulated, if so by what processes/entities? Who is providing the content? [5]. Is this content valid and reliable? Thus, to answer these questions we need to assess trustworthiness of data.

Trust might be created when users want to interchange data. For example, **user1** publishes some data on the web and provide some information to other users. Various users may make use of this information. **User2** receives the provided information. But **user2** wouldn't want to use the available information because it may not be trustworthy enough even if the information is related to his task. Since **user2** wouldn't want to check and verify all information, he decides to consider the information that

[1] Public Key Infrastructure

[2] dereferenceable URI is a resource retrieval mechanism that uses any of the internet protocols (e.g. HTTP) to obtain a copy or representation of the resource it identifies.

originates from trusted providers only as trustworthy information. **User2** can assess the trustworthiness of data by checking its validity and ignore unsuitable objects accordingly. But the question is how a user can check the validity of data? [3].

Regarding this matter, trust over Linked Data can be defined in two parts:

- Authentication of the origin of datasets:
  Trustworthiness and reliability of a dataset should be proved for users and other datasets, so that they can trust the data items. In fact, users need to see the origin of data while retrieving some data piece. So determining ownership for the data and authentication of the ownership is very important.
- Validation of the datasets:
  Besides authentication assessment of Linked Data, many other specializations of the trustworthiness can be considered (e.g. trustworthiness of weblog in a feed aggregator or trustworthiness of photos uploaded to a portal, trustworthiness of statistical reports announced via a news platform).

Another aspect of trust that should be considered is the "user policy". How should a user access to a dataset or RDF document? Furthermore, which parts of an RDF document can be accessible for what users? [6]. Hence, user permission for having access to a document should be regarded in trustworthy of Linked Data.

3. RELATED WORKS

Thuraisingham [9] provided an overview of the secure Semantic Web and considered the security issues for the layered framework of the Semantic Web and described some of the researches on access control and dissemination of XML and RDF documents. Thuraisingham also considered security of information interoperability and providing of security policies on ontologies.

In other direction, the trustworthiness of the Web of Data has been discussed in implementing layers of Linked Data like SPARQL. Hartig [8] described tSPARQL, which is a trust-aware extension to the query language SPARQL. tSPARQL allows us to describe trust requirements in SPARQL queries and an application can filter solutions for graph patterns in SPARQL queries based on the trustworthiness of the data from which the solutions originate. Hartig also proposes a trust model that associates RDF statements with trust values and extends the SPARQL semantics to access to these trust values in tSPARQL and uses a trust function for representation of trustworthiness of each RDF triple and depicts this trustworthy connection among triple by trust weighted RDF graph.

In [15], voidp, a light-weight provenance extension for the voiD[3] vocabulary that allows data publishers to add provenance metadata to the elements of their datasets, has been described. Dataset signature, signature method, certification and authority are some properties that have been considered in the mentioned vocabulary to prove the origin of a dataset and its authentication.

[3] http://semanticweb.org/wiki/VoiD

4. PROPOSED APPROACH

Here the proposed approach is described. First, PKI concepts and its applications are briefly explained. And then we show how to use these concepts for securing of Linked Data. Finally, we propose a way in which dataset can be authorized by a trustworthy center and datasets can be connected together in a trustworthy way using PKI principles.

4.1 PKI Concepts for Securing the Web of Data

PKI is one of the mechanisms that can be used for providing trust in data transfer. The technology is called Public Key because it works with a pair of keys. One of the two keys may be used to encrypt information, which can only be decrypted with the other key. One key is made public and the other one is kept secret. The secret key is usually called the private key [10]. Anyone may obtain pair keys, the public key and the private key. The Infrastructure is the underlying systems needed to issue keys and certificates and to publish the public information. PKI structure assures users of the trust in their documents and objects.

PKI has four important principles: Authentication, Non-repudiation, Encryption and Privacy. As a matter of fact, PKI not only is an authentication method, but also is an infrastructure for issuing a digital certificates and doing some other operations on certificates such as managing and revoking the key pairs that are used to authenticate users and objects within a network or across the web. In fact, a public key certificate is an electronic document which uses a digital signature to bind a public key with identity information such as the name of a person or an organization, their address, and so forth and it can be used to verify that a public key belongs to an individual [11]. Digital certificates certify that the people, the website, and the network resources such as servers and routers are reliable sources, in other words, who or what they claim to be and provide protection for the data exchanged from the visitor and the website from tampering or even theft, such as credit card information. In fact, the purpose of a digital certificate is to reliably link a public/private key pair with its owner.

The third-party who issues certificates is known as a Certification Authority (CA). When a CA issues Digital certificates, it verifies that the owner is not claiming a false identity.

In brief, the goal of a public key infrastructure (PKI) is to enable secure, convenient, and efficient discovery of public keys [12]. It should be applicable within as well as between organizations, and scalable to support web of data.

4.2 Mapping PKI to Linked Data

Here, we discuss a method that exists for active entities (such as people and devices) and map it to the case of the Web of Data. Although datasets may often include information about the data publishers and/or curators, that information might not be accurate. A third party can be used to authenticate the source of data. Originality of an object can be proved by digital signature. In other words, we can use digital signature for authentication of datasets.

When ownership of a digital signature secret key is bound to a specific dataset, a valid signature shows that the data was sent by that dataset. This means non-repudiation of origin or provenance information, is an important aspect of digital signatures. By this property an entity that has signed some information cannot at a later time deny having signed it. Similarly, access to the public key only does not enable a fraudulent party to fake a valid signature.

For identifying a dataset on the web, we should provide a certificate for datasets. A digital certificate for a dataset proves the validity of a dataset and assures users that the dataset is not fake or invalid. It contains some fields such as dataset URI, name, expiration dates, dataset public key (used for encrypting messages), issuer of certificate, and the digital signature of the certificate-issuing authority so that a recipient can verify that the certificate is real.

A trust center (or CA) digitally signs the certificate, thereby attesting to the validity of the certificate's information. There are some policies governing how the CAs issue, manage and revoke certificates and store keys, digital certificates and their keys. Since a user trusts the CA that issues a certificate, and the certificate is valid, the user can trust the certificate.

For instance, an SSL[4] certificate is a kind of certificate that verifies the security and authentication of an interaction. In fact, SSL certificates give a website the ability to communicate securely with its web customers [13]. Without a certificate, any information sent from a client to a website can be intercepted and viewed by hackers and fraudsters. In particular, SSL uses digital certificates that will attest to the binding of a public key to an individual or other entity. They provide verification of the claim that a specific public key does, in fact, belong to the specified entity. As SSL certificate issued by a CA and a trust connection is created between two devices and websites, datasets certificates can assure credibility and authentication of them.

Now let's assume the communication between two datasets. When two datasets connect and establish a trustworthy relationship, they should get their own certificates and pair keys from a CA. When the CA signs certificates of the datasets, they can exchange their data items in a trustworthy way. For instance, if a dataset transfers its RDF data to another dataset securely, the source dataset can encrypt data items (or part of them) by destination dataset's public key and send it to the destination. On the other hand, the destination dataset can decrypt the data items by its own private key. A dataset can fetch other dataset's public key by referring to CA. (Figure 1)

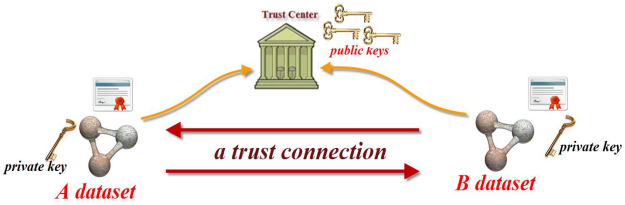

Figure 1- A trust connection between two datasets

Another case can be posed when a dataset (A) does not have any digital certificate and wants to assure another trusted dataset (B). When A send some information to B, sent data and its certificate is checked and verified by destination. Source dataset can also send secure information to B by its public key in a secure way and B can receive encrypted information and decrypt them by its own private key.

With essential concepts of PKI, Linked Data publishers will be able to be supported by reliable sources to publish their data. Having a digital certificate for each dataset (or datasets that we want to protect), a trustworthy relationship can be created among datasets. Trustworthy of web of data will be achieved, if there is a third-party that authenticates objects and datasets. In fact, datasets can be certified by these CAs. Figure 2 depicts how datasets creates trustworthy relationship via a CA and connect them through their public and private keys.

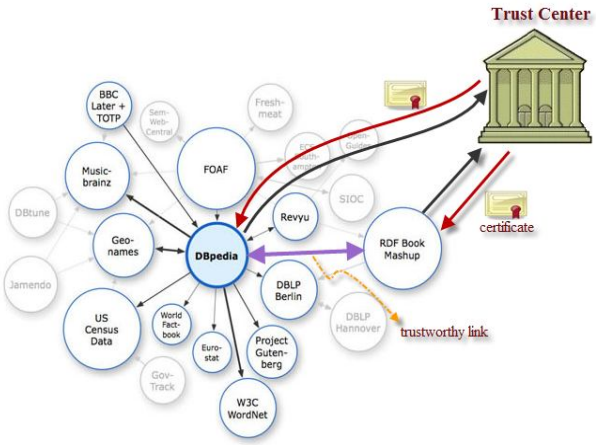

Figure 2- A Trust Center for protecting RDF links

For reaching the goal and authenticating the datasets, each dataset should have a digital signature and some other properties that prove the origin of the dataset. These properties can be included in the metadata of the dataset. Authority and signature are such properties that can be used in the VoiD vocabulary of dataset. Signature property represents the signature of the dataset and authority property shows the authority of the relationship between the item under provenance control and the dataset publisher [15]. If the dataset is signed, a property like certification is used to contain the signature elements and proves the origin of a dataset.
Using Semantic Web Publishing Vocabulary (SWP) [16] as an RDF-Schema vocabulary and voidp[15], we can express information provision related metadata and assure the origin of datasets with digital signatures.

[4] Secure Socket Layer

4.3 Trust Management over Linked Data Using PKI

Since Linked Data users are encouraged to publish their data on the Web and this enables them to have a very powerful and rich database on the Web of Data, there might not be enough sensitivity on trust in Linked Data scope, because we don't want to impose restrictions on Web users. On the other hand, we believe that trustworthy and validity of datasets is very important over the web and without paying special attention to trustworthiness of data, we may encounter a web of data with many invalid links and low quality data.

If datasets are verified by a trust center -we can call it Linked Data CA- in a scope and all datasets accept it as a center of trustworthiness, then users can trust datasets and identify them via their digital certificates. Furthermore, datasets can exchange data in a trustworthy manner. This means they can reach to dereferenceable and valid URIs because their validity is checked by a third-party.

As we mentioned above, one of the methods of datasets authentication is giving them a certificate. The dataset certificate specifies identity and credibility of them. It contains identification information including; the name of a dataset or an organization that publishes the dataset, subject, public key of dataset, number of the contained datasets, validation date of dataset and so forth. This certificate can be used to verify that a public key belongs to a dataset. The signatures on a certificate are attestations by the certificate signer (or Linked Data CA) that indicates the identity information and the public key belong to the same datasets.

Many datasets in Linked Data may have their own CAs and different datasets on different scope can connect each other via their CAs. For example, two CAs can have cross-certificate with each other. A cross-certificate is a certificate issued by one Certificate Authority that signs the public key for the root certificate of another Certificate Authority. Cross-certificates provide the means to create a chain of trust from a single, trusted, root CA to multiple other CAs

5. CONCLUSION

In this paper, we considered trustworthiness of Linked Data and proposed to use PKI principles for Linked Data scope, focusing on authentication and verification of objects and datasets. Having a Linked Data trust center, datasets can get digital certificate from a CA and communicate in a trustworthy manner. Furthermore, they can use all other applications of PKI for exchanging data such as data encryption and privacy.

6. REFERENCES

[1] Halpin, H., A query-driven characterization of linked data. In *Proceedings of the Linked Data Workshop at the World Wide Web Conference*, Madrid, Spain (2009)

[2] Bizer, C., Heath, T. and Berners-Lee, T. 2009. Linked Data — The Story So Far. International. *Journal on Semantic Web and Information Systems,* 2009.

[3] Harting, O., Use Case Simple Trustworthiness Assessment, *http://www.w3.org/2005/Incubator/prov/wiki/Use_Case_Simple_Trustworthiness_Assessment*

[4] Lopez, V., Nikolov, A., Sabou, M., Uren, V., Motta, E. Scaling up question-answering to Linked Data, *17th International Conference on Knowledge Engineering and Knowledge Management*. Lisboa, Portugal, 2010.

[5] Gómez Pérez , J. M., Provenance and Trust, Foundations of Trust in the Future Internet, *Future Internet Assembly, W3C Provenance Incubator Group*, 2010.

[6] Thuraisingham, Bhavani M., Confidentiality, Privacy and Trust Policy Enforcement for the Semantic Web. *POLICY 2007*: 8-11

[7] Thuraisingham, Bhavani M., Building Trustworthy Semantic Webs, *IRI 2009*

[8] Hartig, O., Querying Trust in RDF Data with tSPARQL, In *Proceedings of the 6th European Semantic Web Conference (ESWC)*, Heraklion, Greece, Jun. 2009.

[9] Thuraisingham, Bhavani M., Security Issues for the SemanticWeb, *In Proceedings of the 27th Annual International Computer Software and Applications Conference*, page 632. IEEE, 2003.

[10] http://www.entrust.com/pki.htm

[11] http://en.wikipedia.org/wiki/Public_key_certificate

[12] Perlman, R., An Overview of PKI Trust Models, In *IEEE Network*, vol. 13, 1999, pp. 38-43.

[13] http://www.evsslcertificate.com/ssl/description-ssl.html

[14] Thuraisingham, Bhavani M., Parikh, P., Trustworthy Semantic Web Technologies for Secure Knowledge Management. *EUC (2) 2008*: 186-193

[15] Omitola, T., Zuo, L., Gutteridge, C., Millard, I., Glaser, H., Gibbins, N. and Shadbolt, N. (2011) Tracing the Provenance of Linked Data using voiD. In: *The International Conference on Web Intelligence, Mining and Semantics (WIMS'11)*, May 25 - 27, 2011, Norway.

[16] C. Bizer. , Semantic web publishing vocabulary (swp) user manual, *www4.wiwiss.fu-berlin.de/bizer/WIQA/swp/SWP-UserManual.pdf* (retrieved Nov. 2010), 2006.