Signed metadata: method and application

Emma Tonkin
UKOLN, University of Bath
Tel: +44 1225 38 4930
Fax:+44 1225 38 6838
e.tonkin@ukoln.ac.uk

Julie Allinson
UKOLN, University of Bath
Tel: +44 1225 38 6580
Fax:+44 1225 38 6838
j.allinson@ukoln.ac.uk

As metadata providers increase in number and diversity, and additional contexts for metadata use are identified, issues of trust, provenance and identity gain in relevance. Use of a public-key infrastructure (PKI) is discussed for digital signature of metadata records, providing evidence of the identity of the signer and the authenticity of the information within the record. Two methods are suggested; firstly, the W3C XML-Signature, and secondly, identification of a minimal set of metadata elements that enable signature verification across various character sets and formats, using the OpenPGP standard. Possible strategies for handling annotation within this infrastructure are suggested. Finally, some use cases are briefly discussed.

Keywords:
heterogeneous infrastructure, digital signature, web of trust, provenance.

1. Introduction

The issue of trust, a level of confidence in a source, is of great importance on the Internet in general. The source of a piece of information is a vital detail in analysis; is the source known? Do they generally provide accurate information? Do they have a reason to provide inaccurate information? In this manner, the provenance of a piece of information becomes a necessary detail in analysis and interpretation.

The predominance of the client-server model means that this issue may often be ignored either partially or wholly, particularly in the digital library environment, in that metadata providers are considered to be responsible for the accuracy of their content. Provenance is established either implicitly, or explicitly stated within metadata; the Open Archives Initiative provides the <provenance> tag, permitting versioning of metadata across systems. The model has been further refined in various contexts, such as the DART project [3]. However, the model relies on the accuracy of the metadata provider's data; information could be altered or falsified at any stage in the supply chain. This pattern of trust is possible only because the number of intermediate organisations through which a given metadata record may pass remains relatively low, composed mostly of explicitly trusted organisations, that is, organisations likely to accept the responsibilities implied in provision/use of community resources. Given that the metadata is likely to be potentially of use to developers and end-users outside this community, there is no reason to expect this to remain the case. Indeed, with the growing popularity both of metadata-enabled filesystems and of informal (free-text) metadata tagging services [4], there is reason to expect this information to be reused in

https://doi.org/10.23106/dcmi.952108611

many domains and contexts in whole or in altered form. As distributed architectures become more common, issues of provenance will become both more challenging to resolve and more immediate; as well as the possibility of malformed data, large-scale networks provide greater incentive for abuse, such as spamming, unauthorised data reuse or falsification.
This paper discusses the role that public-key infrastructure (PKI) digital signatures may play in permitting data provenance and accountancy to be handled.

1.1 Principles behind the digital signature

The 'digital signature', first introduced in [1], is an application of cryptographic techniques, that aims to permit the verification of messages. The digital signature operates analogously to a handwritten signature, in that it can be used as proof that the message as received is equal to the message as originally authored. In performing this, a signature also requires as a prerequisite that the identity of the signer is established - to verify that a handwritten signature has not been faked, it must be compared against an existing signature. A similar prerequisite exists in the case of a digital signature. A common use of the digital signature is as an extension of email, to prove the identity of the sender relative to known identities, and to demonstrate the authenticity of the content.
This is possible using a variety of cryptographic methods, but the most common solution today is based around public-key infrastructure (PKI), in which the signer has two keys. One of these, the key used for creating the digital signature, is known as a private key. The other may be distributed to those who need the ability to verify the digital signature, and is known as a public key. The public key cannot be used to sign messages – only for verification. It is therefore possible for any member of the public to verify the signer's identity. This relies on the assumption that the private key remains secret.
Public keys are often distributed by uploading to a *key server*, a repository containing keys and corresponding identity information; some form of public key distribution system is required. It is of course possible for any individual to produce a key with fictitious or stolen identity information, in an attack known as *existential forgery*, which means that this technology is not a general solution to issues of identity – whilst validating a signature as belonging to a known identity is essentially proof of the origin of the signed object, successfully validating a signature as belonging to a 'new' identity reveals only that a person claiming that identity has signed the document. Encountering that signature a second time demonstrates that the same identity generated both documents, in itself a useful piece of information. However, if a key has once been conclusively established to be valid and has not been compromised and revoked, then any documents signed using that key may be trusted as originating from the same, valid individual. A PKI forms a useful role in establishing a network of trust.
Digital signatures have been applied in related domains, such as the Friend of a Friend (FOAF) metadata standard, where the technique is used to solve several problems inherent in the design of FOAF, a decentralised framework for semantic description, by introducing a formal conception of identity [2].

1.2 Approaches to signing DC metadata

An unusual feature of DC metadata is the number of available representations; DC metadata can for example be wrapped in XML, represented in RDF, expressed in

https://doi.org/10.23106/dcmi.952108611

XHTML or stored as plain-text. In the case of XML-formatted data, an appropriate standard for digitally signing the data already exists in the form of the XML Signatures standard, which provides flexible methods for signing and verifying data objects in XML.

There are clear advantages to using a standardised approach, such as a large applicable code base, a standard and well-understood method and an existing development and user base. However, this method is practical only where XML is used, meaning that this approach would not be appropriate for systems employing multiple or alternative metadata representations. Standards such as Z39.50 could not make use of this method in the general case, although certain implementations employ XML as a data format.

One solution is to sign the name-value pairs contained within the record, using an implementation adhering to the OpenPGP standard[1]; these pairs may be used as a basis for authentication such that a change in the encapsulating data format does not invalidate the signature. This presents several difficulties; firstly, since digital signing operates across the byte representation (value) of each character within the string that is to be signed, a change in character encoding such as conversion to UTF-8 would break the signature, as would any change in whitespace characters. Encoding the metadata record in a markup language such as XML/HTML produces a similar effect, provided that the record contains entities with HTML-specific encoding. Since a record does not generally specify its initial (used at the time of signing) character encoding/format, this information must be explicitly appended or established by shared convention. Current character encoding is also required if not given elsewhere, as it is in the case of a valid XHTML document.

The minimal quantity of information required comprises signature method, ID, a pointer to the signed object, and signature. The object to which the signature belongs must be established; in XML Signatures, the signature is wrapped around the XML object to which it belongs. In email, a number of strategies are possible; either the signed data forms a separate MIME attachment, or it is placed between text markers in the body of the email. A shared convention is required in each instance. Unless a shared convention regarding character set and/or encoding is established, these must also be explicitly specified. These issues are handled analogously in XML Signatures [7].

2. Issues in signing metadata

A signed record, by definition, cannot be altered without invalidating the original signature. Editing a record therefore implies that the resulting copy should be re-signed (and potentially re-published) by the editor; the appropriateness of this approach depends on the nature of the communities making use of the metadata. An alternative is to propose changes by means of a feedback mechanism, or to encapsulate changes into 'amendments', a specific form of digital annotation that is to be applied almost as though it were a patch, in the sense of a set of corrections applicable to source code. The establishment of a trust infrastructure permits decisions to be made on a data-source specific level; for example, known-accurate metadata sources may be identified using a whitelist.

A signature forms a container, rather like an envelope; information within the area used for the purpose of generating a signature cannot be edited without tearing the envelope apart. However, information may be appended outside that container. To provide one

1 http://www.openpgp.org

https://doi.org/10.23106/dcmi.952108611

example of a situation in which this may be desirable, certain information to be provided within OAI-DC may be unsuitable for signing into a metadata record; provenance information supplied by an OAI harvester should be signed, not by the institution providing the metadata, who as originators of the metadata itself would sign accordingly, but as an annotation tagged to the record by the OAI harvester. Any model involving digital signatures requires a consensus as to the entity or entities that will take responsibility for each metadata set.

3 Potential applications

A public-key infrastructure forms an additional layer of complexity, resource and infrastructure overhead, and is therefore undesirable outside circumstances in which the functionality is explicitly required or provides clear advantages. A few examples of such situations are briefly discussed here.

3.1 Provenance in aggregation

As the number of repositories and aggregators increases, so too does the number of potential formal or informal metadata sources. The complexity of the process of adding further metadata sources to an aggregator is likely to correspondingly decrease. Currently, trust is conferred informally according to the basis of the perceived reputation and integrity of the source, a solution that scales poorly. The additional provenance information provides a useful aid to solving this problem – if inaccurate or incomplete information is encountered, negative feedback may be collected and remedial action may be taken. For example, 'trust' and 'distrust' ratings may be computed for each provider, permitting search results to be weighted according to expectation of accuracy; a similar approach may be taken for other metrics such as similarity of outlook [5] or methodology of creation.

3.2 A distributed metadata cloud

Though the term 'cloud' has acquired a variety of meanings in other contexts, it is here used in the following sense; a collection of loosely linked nodes, each of which may provide access to data from any of the nodes within the cloud. An aggregator is typically expected to offer end-user services, such as discovery and annotation. In general, such information is neither shared nor passed back to the repository holding the annotated object, nor is it made available to other services or infrastructure elements. However, developers wishing to link an increasing number of heterogeneous metadata sources and services are likely to treat both repository and aggregator of formal metadata as possible sources amongst many; metadata and annotations may be transmitted, stored and made available by otherwise unrelated third-party services.

In a distributed environment, strict marking of provenance and identity may fulfil a number of functions, including access to the metrics and trust mechanism previously described, as well as provision of a data point from which to handle implicit information useful for analysis, such as the record's composition, underlying application profile and local convention. The origin of annotations may be verifiably recorded, as may each stage of the record's transmission path.

3.3 Metadata handling and trust in mobile devices and ad hoc networks

The use of lightweight PKI in mobile devices has previously been explored elsewhere [6]. Due to power and resource limitations in this context, a different choice of PKI may

https://doi.org/10.23106/dcmi.952108611

be required. Since no centralised services are available, ad hoc networks severely limit the use of an implicit trust model; therefore, trust must be modeled on a strictly local level. All interactions take the form of gossip; *I heard from A that B said that C...*
Provided that an appropriate solution is applied to the problem of key distribution, PKI provides methods by which each stage in the chain may be documented. To demonstrate the practical use of such a method: the local transport authority provides a bus timetable in PDF format, described using an appropriately signed metadata record. This record is shared across the network. A user who reads it checks its validity according to the local authority's public key; knowing that the identity is valid, and that the information is as originally published, she decides to trust the provided description, downloading the file from the address provided in the record.

Conclusion

Large-scale and distributed networks may suffer from misuse, such as the injection of false or inaccurate data; examples of 'metadata spammers' can already be seen on tagging services such as del.icio.us. Public-key infrastructure functionality provides methods by which provenance may be demonstrated and defensive steps may be taken, such as the establishment of networks of trust.

References

1. W. Diffie and M. E. Hellman. New Directions in Cryptography. In *IEEE Transactions on Information Theory.* Vol IT-22, No. 6. November 1976.
2. E. Dumbill. Support online communities with FOAF: How the friend-of-a-friend vocabulary addresses issues of accountability and privacy. In IBM's XML Watch, August 2002.
3. G. Dahlquist, B. Hoffman and D. Millman, "Integrating digital libraries and electronic publishing in the DART project", *Proceedings of the fifth ACM/IEEE-CS Joint Conference on Digital Libraries,* 2005, pp. 114-120.
4. T. Hammond, T. Hannay, B. Lund, J. Scott. Social Bookmarking Tools (I): A General Review. In: *D-Lib Magazine* 11 (4), April 2005.
5. J. Golbeck. Semantic Web Interaction through Trust Network Recommender Systems. In: *End User Semantic Web Interaction Workshop* at the 4th International Semantic Web Conference, November 2005.
6. N Smart and H Muller. A wearable public key infrastructure (WPKI). In: *Proceedings IEEE International Symposium on Wearable Computers*, Blair MacIntyre and Bob Iannucci, editors, pages 127--133. IEEE Computer Society, October 2000.
7. Example of an XML Signature. http://java.sun.com/webservices/docs/1.4/tutorial/doc/XMLDigitalSignatureAPI7.html. Retrieved May 2006

https://doi.org/10.23106/dcmi.952108611