Signed metadata : method and application

International Conference on Dublin Core and Metadata Applications, 3 – 6 October 2006, Mexico

Julie Allinson (presenter)
Repositories Research Officer
UKOLN, University of Bath

Emma Tonkin (author)
Interoperability Focus Officer
UKOLN, University of Bath

UKOLN is supported by: JISC, Supported by MLA Museums, Libraries and Archives Council

Contents

- Overview
- Introduction
- Brief background on digital signatures
- Signing DC metadata – approaches and issues
- Use cases

Overview

- Why do we need to digitally sign metadata records?

  Currently, we (probably) don't, but …

  increasing numbers of metadata providers

  + additional ways of reusing data

  = increasing issues of trust, provenance and identity

- Digitally signing metadata records through a Public-Key Infrastructure (PKI) is one potential solution

Introduction

- The current digital library world works on
  - Implicit trust - metadata providers are trusted because we 'know' them
  - Explicit trust – e.g. the OAI-PMH <provenance> tag provides information
  - A small network of trusted and responsible organisations
- In the future, we can envisage
  - Metadata-enabled filesystems
  - Increased informal metadata tagging services
  - Larger-scale networks
  - More opportunities for abuse, e.g. spamming
  - Less accountability and responsibility (= less trust)

Digital signatures

- Date back to 1976
- Use cryptographic techniques
- Similar to handwritten signatures
- Permit the verification of messages

The most common solution is Public-Key Infrastructure (PKI)

PKI – how does it work?

- A digital signer has 2 keys
  - Private key – used to create the signature
  - Public key – used by third-parties to verify the author
- Public keys are distributed by a distribution system, e.g. a key server containing keys and identity information
- PKI is useful in establishing a network of trust
- … but it has limitations
  - It is possible to produce a key with fictitious, false or stolen identity

Signing Dublin Core metadata

- Dublin Core is unusual in that it can be represented in different ways, e.g. XML, RDF, XHTML
- Approaches
  - The XML Signatures standard provides flexible methods for signing and verifying data objects in XML
    - An XML metadata record could be wrapped within an XML Signature
  - OpenPGP is an alternative mechanism
    - OpenPGP could be used to sign the name-value pairs within a metadata record
- A standardised approach is required for both mechanisms

XML Signatures (1)

```
<Signature ID?>
        <SignedInfo>
          <CanonicalizationMethod/>
          <SignatureMethod/>
          (<Reference URI? >
            (<Transforms>)?
            <DigestMethod>
            <DigestValue>
          </Reference>)+
        </SignedInfo>
        <SignatureValue>
       (<KeyInfo>)?
       (<Object ID?>)*
    </Signature>
```

"XML digital signatures are represented by the Signature element which has the following structure (where "?" denotes zero or one occurrence; "+" denotes one or more occurrences; and "*" denotes zero or more occurrences)"

(from http://www.w3.org/TR/xmldsig-core/)

XML Signatures (2)

- Reference URI='
  http://example.com/the-signed-dc-record'>
  <DigestMethod/> (the type of signature used)
  <DigestValue></DigestValue> (contains the signature, an encrypted value)

```
<metadata
      xmlns="http://example.org/myapp/"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="http://example.org/myapp/ http://example.org/myapp/schema.
    xsd"
      xmlns:dc="http://purl.org/dc/elements/1.1/">
     <dc:title>
       UKOLN
      </dc:title>
      <dc:description>
       ...
      </dc:description>
      <dc:publisher>
       UKOLN, University of Bath
      </dc:publisher>
      <dc:identifier>
       http://www.ukoln.ac.uk/
      </dc:identifier>
    </metadata>
    <dsig:Signature
        xmlns:dsig="http://www.w3.org/2000/09/xmldsig#"
        xmlns:dsig-xpath="http://www.w3.org/2002/06/xmldsig-filter2">
         <dsig:SignedInfo>
           ...
         <dsig:Reference URI="#metadata">
    </dsig:Signature>
```

(see http://dublincore.org/documents/dc-xml-guidelines/)

Minimum components of a signature

- XML Signatures is not the only method for signing metadata. The following information would be needed for any signature:
  - character set, encoding
  - current character encoding (to enable conversion)
  - signature method (e.g. sha1)
  - ID - the ID of the signer, could be an email address
  - the signature itself, analogously referred to in XML-Signature as 'digestvalue‘
  - Information about the signed item, such as the metadata schema
    - for example, use of 'oai-dc‘ might be taken to mean 'expect all of the oai-dc elements to be present in key-value pairs, check the signature over all of them'

Issues in signing metadata

- Changes in encoding and/or character set will invalidate signatures
- A signature is also invalidated if changes are made to the metadata record. The ‘new’ package should be re-signed by whoever makes those changes and the original signed package is effectively lost.
- How do we maintain the integrity of the original signature and the original metadata record?
- Alternative methods
  - digital amendments or annotations appended outside of the original package
- Information supplied by an OAI harvester might be signed by the OAI harvester (amendments/annotations) and/or by the repository (unchanged metadata).
  - Undersigning all metadata by the harvester offers a kind of ‘traceroute’ to show the history of that record
  - But it could lead to large packets of metadata being transferred around networks

Provenance in aggregation

- Currently
  - aggregators are most likely to harvest content from the originating repository
- In the future the repository ecology looks much more complex
  - increased repository numbers and sharing metadata between repositories
  - more aggregators and aggregation of content from more sources
  - increased availability of informal metadata sources
- Current trust mechanisms (perceived integrity of the source) do not scale
- PKI could be used to identify the origination of the metadata and its route through other repositories and/or aggregators

Potential applications

- A public-key infrastructure adds complexity, resource and infrastructure overheads
- It is valuable only where the functionality is explicitly required or provides clear advantages
- Some examples
  - Provenance in aggregation
  - A distributed metadata cloud
  - Metadata handling and trust in mobile devices and ad-hoc networks

Distributed metadata cloud

- = a loosely coupled, interoperating collection of heterogeneous metadata sources and other services
- Information is seamlessly passed between members of the ‘cloud’
- Identifying provenance and identity provides
  - A trust mechanism for assessing the potential value of information
  - A verifiable transmission path and origin of annotations
  - Access to additional information about the data source

Mobile devices and ad hoc networks

- In a centralised system it is relatively easy to ascertain the originator of information
- … but with increasingly pervasive ad hoc Internet access
- offered in a decentralised way
- lightweight PKI can help identify each stage in the chain and thereby help us distinguish the spam from the trusted

Conclusion

- Issues of provenance and identity are dealt with in the current digital library realm by the perceived integrity of a source
- As the number of metadata sources and aggregators increase, these informal mechanisms may prove insufficient and metadata may be subject to abuse
- Digitally signing metadata records can help to identify provenance
- Public key infrastructure functionality offers particular cryptographic methods to digitally signing metadata
- And can help to create new networks of trust