Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: November 10, 2017

P. Hallam-Baker
Comodo Group Inc.

May 9, 2017

# Uniform Data Fingerprint (UDF)
# draft-hallambaker-udf-05

## Abstract

This document describes means of generating Uniform Data Fingerprint (UDF) values and their presentation as text sequences and as URIs.

Cryptographic digests provide a means of uniquely identifying static data without the need for a registration authority. A fingerprint is a form of presenting a cryptographic digest that makes it suitable for use in applications where human readability is required. The UDF fingerprint format improves over existing formats through the introduction of a compact algorithm identifier affording an intentionally limited choice of digest algorithm and the inclusion of an IANA registered MIME Content-Type identifier within the scope of the digest input to allow the use of a single fingerprint format in multiple application domains.

Alternative means of rendering fingerprint values are considered including machine-readable codes, word and image lists.

## Status of This Memo

Copyright Notice



Table of Contents

## 1. Definitions

Cryptographic Digest Function

A hash function that has the properties required for use as a cryptographic hash function. These include collision resistance, first pre-image resistance and second pre-image resistance.

Content Type

An identifier indicating how a Data Value is to be interpreted as specified in the IANA registry Media Types.

Data Value

The binary octet stream that is the input to the digest function used to calculate a digest value.

Data Object

A Data Value and its associated Content Type

Digest Algorithm

A synonym for Cryptographic Digest Function

Digest Value

The output of a Cryptographic Digest Function

Data Digest Value

The output of a Cryptographic Digest Function for a given Data Value input.

Fingerprint

A presentation of the digest value of a data value or data object.

Fingerprint Presentation

The representation of at least some part of a fingerprint value in human or machine readable form.

Fingerprint Improvement

The practice of recording a higher precision presentation of a fingerprint on successful validation.

Fingerprint Work Hardening

The practice of generating a sequence of fingerprints until one is found that matches criteria that permit a compressed presentation form to be used. The compressed fingerprint thus being shorter than but presenting the same work factor as an uncompressed one.

Hash

A function which takes an input and returns a fixed-size output. Ideally, the output of a hash function is unbiased and not correlated to the outputs returned to similar inputs in any predictable fashion.

Precision

The number of significant bits provided by a Fingerprint Presentation.

Work Factor

A measure of the computational effort required to perform an attack against some security property.

## 1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

## 2. Introduction

The use of cryptographic digest functions to produce identifiers is well established as a means of generating a unique identifier for fixed data without the need for a registration authority.

While the use of fingerprints of public keys was popularized by PGP, they are employed in many other applications including OpenPGP, SSH, BitCoin and PKIX.

A cryptographic digest is a particular form of hash function that has the properties:

It is easy to compute the digest value for any given message

It is infeasible to generate a message from its digest value

It is infeasible to modify a message without changing the digest value

It is infeasible to find two different messages with the same digest value.

If these properties are met, the only way that two data objects that map to the same digest value is by random chance. If the number of possible digest values is sufficiently large (i.e. is a sufficiently large number of bits in length), this chance is reduced to an arbitrarily infinitesimal probability. Such values are described as being probabilistically unique.

A fingerprint is a representation of a cryptographic digest value optimized for purposes of verification and in some cases data entry.

## 2.1. Algorithm Identifier

Although a secure cryptographic digest algorithm has properties that make it ideal for certain types of identifier use, several cryptographic digest algorithms have found widespread use, some of which have been demonstrated to be insecure.

For example the MD5 message digest algorithm [RFC1321], was widely used in IETF protocols until it was demonstrated to be vulnerable to collision attacks [TBS].

The secure use of a fingerprint scheme therefore requires the digest algorithm to either be fixed or otherwise determined by the fingerprint value itself. Otherwise an attacker may be able to use a weak, broken digest algorithm to generate a data object matching a fingerprint value generated using a strong digest algorithm.

## 2.2. Content Type Identifier

A secure cryptographic digest algorithm provides a unique digest value that is probabilistically unique for a particular byte sequence but does not fix the context in which a byte sequence is interpreted. While such ambiguity may be tolerated in a fingerprint format designed for a single specific field of use, it is not acceptable in a general purpose format.

For example, the SSH and OpenPGP applications both make use of fingerprints as identifiers for the public keys used but using different digest algorithms and data formats for representing the public key data. While no such vulnerability has been demonstrated to date, it is certainly conceivable that a crafty attacker might construct an SSH key in such a fashion that OpenPGP interprets the data in an insecure fashion. If the number of applications making use of fingerprint format that permits such substitutions is sufficiently large, the probability of a semantic substitution vulnerability being possible becomes unacceptably large.

A simple control that defeats such attacks is to incorporate a content type identifier within the scope of the data input to the hash function.

## 2.3. Representation

The representation of a fingerprint is the format in which it is presented to either an application or the user.

Base32 encoding is used to produce the preferred text representation of a UDF fingerprint. This encoding uses only the letters of the Latin alphabet with numbers chosen to minimize the risk of ambiguity between numbers and letters (2, 3, 4, 5, 6 and 7).

To enhance readability and improve data entry, characters are grouped into groups of five.

## 2.4. Truncation

Different applications of fingerprints demand different tradeoffs between compactness of the representation and the number of significant bits. A larger the number of significant bits reduces the risk of collision but at a cost to convenience.

Modern cryptographic digest functions such as SHA-2 produce output values of at least 256 bits in length. This is considerably larger than most uses of fingerprints require and certainly greater than can be represented in human readable form on a business card.

Since a strong cryptographic digest function produces an output value in which every bit in the input value affects every bit in the output value with equal probability, it follows that truncating the digest value to produce a finger print is at least as strong as any other mechanism if digest algorithm used is strong.

Using truncation to reduce the precision of the digest function has the advantage that a lower precision fingerprint of some data content

is always a prefix of a higher prefix of the same content. This allows higher precision fingerprints to be converted to a lower precision without the need for special tools.

## 3. Encoding

A UDF fingerprint for a given data object is generated by calculating the Binary Fingerprint Value for the given data object and type identifier, truncating it to obtain the desired degree of precision and then converting the truncated value to a representation.

### 3.1. Binary Fingerprint Value

The binary encoding of a fingerprint is calculated using the formula:

Fingerprint = <<Version-ID> + H (<<Content-ID> + ?:? + H(<<Data>))

Where

H(x) is the cryptographic digest function
<<Version-ID> is the fingerprint version and algorithm identifier.
<<Content-ID> is the MIME Content-Type of the data.
<<Data> is the binary data.

The use of the nested hash function permits a fingerprint to be taken of data for which a digest value is already known without the need to calculate a new digest over the data.

The inclusion of a MIME content type prevents message substitution attacks in which one content type is substituted for another.

#### 3.1.1. Version ID

A Version Identifier consists of a single byte. The following digest algorithm identifiers are specified in this document:

SHA-2-512 = 96

SHA-2-512 (compressed) = 97, 98, 99, 100

SHA-3-512 = 144

These algorithm identifiers have been chosen so that the first character in a SHA-2-512 fingerprint will always be 'M' and the first character in a SHA-3-512 fingerprint will always be 'S'. These provide mnemonics for 'Merkle-Damgard' and 'Sponge' respectively.

### 3.2. Truncation

The Binary Fingerprint Value is truncated to an integer multiple of 25 bits regardless of the intended output presentation.

The output of the hash function is truncated to a sequence of n bits by first selecting the first n/8 bytes of the output function. If n is an integer multiple of 8, no additional bits are required and this is the result. Otherwise the remaining bits are taken from the most significant bits of the next byte and any unused bits set to 0.

For example, to truncate the byte sequence [a0, b1, c2, d3, e4] to 25 bits. 25/8 = 3 bytes with 1 bit remaining, the first three bytes of the truncated sequence is [a0, b1, c2] and the final byte is e4 AND 80 = 80 which we add to the previous result to obtain the final truncated sequence of [a0, b1, c2, 80]

### 3.3. Base32 Representation

A modified version of Base32 [RFC4648] encoding is used to present the fingerprint in text form grouping the output text into groups of five characters separated by a dash '-'. This representation improves the accuracy of both data entry and verification.

### 3.4. Examples

In the following examples, <Content-ID> is the UTF8 encoding of the string "text/plain" and is the UTF8 encoding of the string "UDF Data Value"

Data = 55 44 46 20 44 61 74 61 20 56 61 6c 75 65

#### 3.4.1. Using SHA-2-512 Digest

```
H( <Data> ) =
    48 da 47 cc  ab fe a4 5c  76 61 d3 21  ba 34 3e 58
    10 87 2a 03  b4 02 9d ab  84 7c ce d2  22 b6 9c ab
    02 38 d4 e9  1e 2f 6b 36  a0 9e ed 11  09 8a ea ac
    99 d9 e0 bd  ea 47 93 15  bd 7a e9 e1  2e ad c4 15
H(H( <Data> ) + Content-ID>) =
    45 e0 59 e0  39 34 ea b7  f6 5d 83 b2  d8 f9 b1 6d
    2a 6b 08 63  d9 3c c1 02  86 7b 83 49  f2 d9 f0 8f
    fe 07 87 30  c7 c9 05 74  ac a1 38 2b  b3 14 4d c6
    39 f9 8c 12  c0 4a 3e b5  05 0b 3e 67  df 52 4b 57
```

Text Presentation (100 bit)MB2GK-6DUF5-YGYYL-JNY5E

Text Presentation (125 bit)MB2GK-6DUF5-YGYYL-JNY5E-RWSHZ

Text Presentation (150bit)MB2GK-6DUF5-YGYYL-JNY5E-RWSHZ-SV75J

Text Presentation (250bit)MB2GK-6DUF5-YGYYL-JNY5E-RWSHZ-SV75J-C4OZQ-5GIN2-GQ7FQ-EEHFI

### 3.5. Fingerprint Improvement

Since an application must always calculate the full fingerprint value as part of the verification process, an application MAY record a

Applications are encouraged to make use of the practice of fingerprint improvement wherever possible.

### 3.6. Compressed Presentation

Fingerprint compression permits the use of shorter fingerprint presentation without a reduction in the attacker work factor by requiring the fingerprint value to match a particular pattern.

UDF fingerprints MUST use compression if possible. A compressed fingerprint uses a version identifier that specifies the form of compression used as follows:

96 No compression

97 First 25 bits are zeros

98 First 40 bits are zeros

99 First 50 bits are zeros

100 First 55 bits are zeros

Thus the fingerprint that would be represented in uncompressed form as MAAAA-AAWIY-LTMFTG-CZTRO is instead represented as MIWIY-LTMFTG-CZTRO.

### 3.7. Identifiers formed using UDFs

UDF fingerprints MAY be used to form a part of another protocol identifier. Such practice carries the implicit semantic that the interpretation of the identifier formed is bound to the document identified by the fingerprint.

### 3.7.1. URI Representation

Any UDF fingerprint MAY be encoded as a URI by prefixing the Base32 text representation of the fingerprint with the string 'udf:'

### 3.7.2. DNS Name

A UDF fingerprint MAY be encoded as a DNS label by prefixing the Base32 text representation with the string 'zz--'.

A DNS name that includes a UDF fingerprint as a DNS label carries the implicit assertion that the interpretation of the address MUST be authorized by a security policy that is validated under a key that matches the corresponding fingerprint.

Placing such a DNS label as the top level (rightmost) label in a DNS address creates an address that is not legal and thus cannot be resolved by the Internet DNS infrastructure. Thus ensuring that the address is rejected by applications that are not capable of performing the associated validation steps.

For example, Alice has the email security key with fingerprint MB2GK-6DUF5-YGYYL-JNY5E. She uses the following email addresses:

alice@example.com

Alice publishes this email address when she does not want the other party to use the secure email system.

alice@zz--mb2gk-6duf5-ygyyl-jny5e.example.com

Alice publishes this email address when she wants to give the other party the option of using secure email if their system supports it.

The DNS server for example.com has been configured to redirect requests to resolve zz--mb2gk-6duf5-ygyyl-jny5e.example.com to the mail server example.com.

alice@example.com.zz--mb2gk-6duf5-ygyyl-jny5e

Alice uses this email address when she wants the other party to be able to send her email if and only if their client supports use of the secure messaging system.

While there should never be a DNS label of the form zz--* in the authoritative DNS root, such labels MAY be introduced by a trusted local resolver. This would allow attempts at making an untrusted

communication request to be transparently redirected through a locally trusted security enhancing proxy.

## 4. Content Types

While a UDF fingerprint MAY be used to identify any form of static data, the use of a UDF fingerprint to identify a public key signature key provides a level of indirection and thus the ability to identify dynamic data. The content types used to identify public keys are thus of particular interest.

As described in the security considerations section, the use of fingerprints to identify a bare public key and the use of fingerprints to identify a public key and associated security policy information are very different.

### 4.1. PKIX Certificates and Keys

UDF fingerprints MAY be used to identify PKIX certificates, CRLs and public keys in the ASN.1 encoding used in PKIX certificates.

Since PKIX certificates and CLRs contain security policy information, UDF fingerprints used to identify certificates or CRLs SHOULD be presented with a minimum of 200 bits of precision. PKIX applications MUST not accept UDF fingerprints specified with less than 200 bits of precision for purposes of identifying trust anchors.

PKIX certificates, keys and related content data are identified by the following content types:

application/pkix-cert

A PKIX Certificate

application/pkix-crl

A PKIX CRL

application/pkix-keyinfo

The KeyInfo structure defined in the PKIX certificate specification

### 4.2. OpenPGP Key

OpenPGPv5 keys and key set content data are identified by the following content types:

application/pgp-key-v5

An OpenPGP key

application/pgp-keys

An OpenPGP key set.

### 4.3. DNSSEC

DNSSEC record data consists of DNS records which are identified by the following content type:

application/dns

A DNS resource record in binary format

## 5. Additional UDF Renderings

By default, a UDF fingerprint is rendered in the Base32 encoding described in this document. Additional renderings MAY be employed to facilitate entry and/or verification of fingerprint values.

### 5.1. Machine Readable Rendering

The use of a machine-readable rendering such as a QR Code allows a UDF value to be input directly using a smartphone or other device equipped with a camera.

A QR code fixed to a network capable device might contain the fingerprint of a machine readable description of the device.

### 5.2. Word Lists

The use of a Word List to encode fingerprint values was introduced by Patrick Juola and Philip Zimmerman for the PGPfone application. The PGP Word List is designed to facilitate exchange and verification of fingerprint values in a voice application. To minimize the risk of misinterpretation, two word lists of 256 values each are used to encode alternative fingerprint bytes. The compact size of the lists used allowed the compilers to curate them so as to maximize the phonetic distance of the words selected.

The PGP Word List is designed to achieve a balance between ease of entry and verification. Applications where only verification is required may be better served by a much larger word list, permitting shorter fingerprint encodings.

For example, a word list with 16384 entries permits 14 bits of the fingerprint to be encoded at once, 65536 entries permits 16. These

encodings allow a 125 bit fingerprint to be encoded in 9 and 8 words respectively.

### 5.3. Image List

An image list is used in the same manner as a word list affording rapid visual verification of a fingerprint value. For obvious reasons, this approach is not generally suited to data entry.

## 6. Security Considerations

### 6.1. Work Factor and Precision

A given UDF data object has a single fingerprint value that may be presented at different precisions. The shortest legitimate precision with which a UDF fingerprint may be presented has 96 significant bits

A UDF fingerprint presents the same work factor as any other cryptographic digest function. The difficulty of finding a second data item that matches a given fingerprint is 2^n and the difficulty or finding two data items that have the same fingerprint is 2^(n/2). Where n is the precision of the fingerprint.

For the algorithms specified in this document, n = 512 and thus the work factor for finding collisions is 2^256, a value that is generally considered to be computationally infeasible.

Since the use of 512 bit fingerprints is impractical in the type of applications where fingerprints are generally used, truncation is a practical necessity. The longer a fingerprint is, the less likely it is that a user will check every character. It is therefore important to consider carefully whether the security of an application depends on second pre-image resistance or collision resistance.

In most fingerprint applications, such as the use of fingerprints to identify public keys, the fact that a malicious party might generate two keys that have the same fingerprint value is a minor concern. Combined with a flawed protocol architecture, such a vulnerability may permit an attacker to construct a document such that the signature will be accepted as valid by some parties but not by others.

For example, Alice generates keypairs until two are generated that have the same 100 bit UDF presentation (typically 2^48 attempts). She registers one keypair with a merchant and the other with her bank. This allows Alice to create a payment instrument that will be accepted as valid by one and rejected by the other.

The ability to generate of two PKIX certificates with the same fingerprint and different certificate attributes raises very different and more serious security concerns. For example, an attacker might generate two certificates with the same key and different use constraints. This might allow an attacker to present a highly constrained certificate that does not present a security risk to an application for purposes of gaining approval and an unconstrained certificate to request a malicious action.

In general, any use of fingerprints to identify data that has security policy semantics requires the risk of collision attacks to be considered. For this reason the use of short, 'user friendly' fingerprint presentations (Less than 200 bits) SHOULD only be used for public key values.

### 6.2. Semantic Substitution

Many applications record the fact that a data item is trusted, rather fewer record the circumstances in which the data item is trusted. This results in a semantic substitution vulnerability which an attacker may exploit by presenting the trusted data item in the wrong context.

The UDF format provides protection against high level semantic substitution attacks by incorporating the content type into the input to the outermost fingerprint digest function. The work factor for generating a UDF fingerprint that is valid in both contexts is thus the same as the work factor for finding a second preimage in the digest function (2^512 for the specified digest algorithms).

It is thus infeasible to generate a data item such that some applications will interpret it as a PKIX key and others will accept as an OpenPGP key. While attempting to parse a PKIX key as an OpenPGP key is virtually certain to fail to return the correct key parameters it cannot be assumed that the attempt is guaranteed to fail with an error message.

The UDF format does not provide protection against semantic substitution attacks that do not affect the content type.

## 7. IANA Considerations

[This will be extended later]

### 7.1. URI Registration

[Here a URI registration for the udf: scheme]

### 7.2. Content Type Registration

[application/pkix-keyinfo]

[application/pgp-key]

### 7.3. Version Registry

96 = SHA-2-512

97 = SHA-2-512 with 25 leading zeros

98 = SHA-2-512 with 40 leading zeros

99 = SHA-2-512 with 50 leading zeros

100 = SHA-2-512 with 55 leading zeros

144 = SHA-3-512

## 8. Normative References

[RFC1321] Rivest, R., "The MD5 Message-Digest Algorithm", RFC 1321, DOI 10.17487/RFC1321, April 1992.

[RFC4648] Josefsson, S., "The Base16, Base32, and Base64 Data Encodings", RFC 4648, DOI 10.17487/RFC4648, October 2006.

Author's Address

Phillip Hallam-Baker
Comodo Group Inc.

Email: philliph@comodo.com