Internet Engineering Task Force (IETF)
Request for Comments: 6885
Category: Informational
ISSN: 2070-1721

M. Blanchet
Viagenie
A. Sullivan
Dyn, Inc.

March 2013

Stringprep Revision and Problem Statement for the Preparation and Comparison of Internationalized Strings (PRECIS)

Abstract

If a protocol expects to compare two strings and is prepared only for those strings to be ASCII, then using Unicode code points in those strings requires they be prepared somehow. Internationalizing Domain Names in Applications (here called IDNA2003) defined and used Stringprep and Nameprep. Other protocols subsequently defined Stringprep profiles. A new approach different from Stringprep and Nameprep is used for a revision of IDNA2003 (called IDNA2008). Other Stringprep profiles need to be similarly updated, or a replacement of Stringprep needs to be designed. This document outlines the issues to be faced by those designing a Stringprep replacement.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for informational purposes.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are a candidate for any level of Internet Standard; see Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc6885.

Table of Contents

1. Introduction

Internationalizing Domain Names in Applications (here called IDNA2003) [RFC3490] [RFC3491] [RFC3492] and [RFC3454] describes a mechanism for encoding Unicode labels that make up the Internationalized Domain Names (IDNs) as standard DNS labels. The labels were processed using a method called Nameprep [RFC3491] and Punycode [RFC3492]. That method was specific to IDNA2003 but is generalized as Stringprep [RFC3454]. The general mechanism is used by other protocols with similar needs but with different constraints than IDNA2003.

Stringprep defines a framework within which protocols define their Stringprep profiles. Some known IETF specifications using Stringprep are listed below:

- The Nameprep profile [RFC3490] for use in Internationalized Domain Names (IDNs);
- The Inter-Asterisk eXchange (IAX) using Nameprep [RFC5456];
- NFSv4 [RFC3530] and NFSv4.1 [RFC5661];
- The Internet Small Computer System Interface (iSCSI) profile [RFC3722] for use in iSCSI names;
- The Extensible Authentication Protocol (EAP) [RFC3748];
- The Nodeprep and Resourceprep profiles [RFC3920] (which was obsoleted by [RFC6120]) for use in the Extensible Messaging and Presence Protocol (XMPP), and the XMPP to Common Presence and Instant Messaging (CPIM) mapping [RFC3922] (the latter of these relies on the former);
- The Internationalized Resource Identifier (IRI) and URI in XMPP [RFC5122];
- The Policy MIB profile [RFC4011] for use in the Simple Network Management Protocol (SNMP);
- Transport Layer Security (TLS) [RFC4279];
- The Lightweight Directory Access Protocol (LDAP) profile [RFC4518] for use with LDAP [RFC4511] and its authentication methods [RFC4513];
- PKIX subject identification using LDAPprep [RFC4683];

- PKIX Certificate Revocation List (CRL) using LDAPprep [RFC5280];
- The Simple Authentication and Security Layer (SASL) [RFC4422] and SASLprep profile [RFC4013] for use in SASL;
- Plain SASL using SASLprep [RFC4616];
- SMTP Auth using SASLprep [RFC4954];
- The Post Office Protocol (POP3) Auth using SASLprep [RFC5034];
- TLS Secure Remote Password (SRP) using SASLprep [RFC5054];
- SASL Salted Challenge Response Authentication Mechanism (SCRAM) using SASLprep [RFC5802];
- Remote management of Sieve using SASLprep [RFC5804];
- The Network News Transfer Protocol (NNTP) using SASLprep [RFC4643];
- IMAP4 using SASLprep [RFC4314];
- The trace profile [RFC4505] for use with the SASL ANONYMOUS mechanism;
- Internet Application Protocol Collation Registry [RFC4790];
- The unicode-casemap Unicode Collation [RFC5051].

However, a review (see [78PRECIS]) of these protocol specifications found that they are very similar and can be grouped into a short number of classes. Moreover, many reuse the same Stringprep profile, such as the SASL one.

IDNA2003 was replaced because of some limitations described in [RFC4690]. The new IDN specification, called IDNA2008 [RFC5890], [RFC5891], [RFC5892], [RFC5893] was designed based on the considerations found in [RFC5894]. One of the effects of IDNA2008 is that Nameprep and Stringprep are not used at all. Instead, an algorithm based on Unicode properties of code points is defined. That algorithm generates a stable and complete table of the supported Unicode code points for each Unicode version. This algorithm uses an inclusion-based approach, instead of the exclusion-based approach of Stringprep/Nameprep. That is, IDNA2003 created an explicit list of excluded or mapped-away characters; anything in Unicode 3.2 that was not so listed could be assumed to be allowed under the protocol.

IDNA2008 begins instead from the assumption that code points are disallowed and then relies on Unicode properties to derive whether a given code point actually is allowed in the protocol.

This document lists the shortcomings and issues found by protocols listed above that defined Stringprep profiles. It also lists the requirements for any potential replacement of Stringprep.

2. Keywords

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 [RFC2119].

This document uses various internationalization terms, which are defined and discussed in [RFC6365].

Additionally, this document defines the following keyword:

PRECIS: Preparation and Comparison of Internationalized Strings

3. Conventions

A single Unicode code point in this memo is denoted by "U+" followed by four to six hexadecimal digits, as used in [Unicode61], Appendix A.

4. Stringprep Profiles Limitations

During IETF 77 (March 2010), a BOF discussed the current state of the protocols that have defined Stringprep profiles [NEWPREP]. The main conclusions from that discussion were as follows:

- Stringprep is bound to Version 3.2 of Unicode. Stringprep has not been updated to new versions of Unicode. Therefore, the protocols using Stringprep are stuck at Unicode 3.2, and their specifications need to be updated to support new versions of Unicode.

- The protocols would like to not be bound to a specific version of Unicode, but rather have better Unicode version agility in the way of IDNA2008. This is important partly because it is usually impossible for an application to require Unicode 3.2; the application gets whatever version of Unicode is available on the host.

- The protocols require better bidirectional support (bidi) than currently offered by Stringprep.

- If the protocols are updated to use a new version of Stringprep or another framework, then backward compatibility is an important requirement. For example, Stringprep normalization is based on and profiles may use Unicode Normalization Form KC (NFKC) [UAX15], while IDNA2008 mostly uses Unicode Normalization Form C (NFC) [UAX15].

- Identifiers are passed between protocols. For example, the same username string of code points may be passed between SASL, XMPP, LDAP, and EAP. Therefore, a common set of rules or classes of strings are preferred over specific rules for each protocol. Without real planning in advance, many Stringprep profiles reuse other profiles, so this goal was accomplished by accident with Stringprep.

Protocols that use Stringprep profiles use strings for different purposes:

- XMPP uses a different Stringprep profile for each part of the XMPP address Jabber Identifier (JID): a localpart, which is similar to a username and used for authentication; a domainpart, which is a domain name; and a resourcepart, which is less restrictive than the localpart.

- iSCSI uses a Stringprep profile for the names of protocol participants (called initiators and targets). The iSCSI Qualified Name (IQN) format of iSCSI names contains a reversed DNS domain name.

- SASL and LDAP use a Stringprep profile for usernames.

- LDAP uses a set of Stringprep profiles.

The apparent judgement of the BOF attendees [NEWPREP] was that it would be highly desirable to have a replacement of Stringprep, with similar characteristics to IDNA2008. That replacement should be defined so that the protocols could use internationalized strings without a lot of specialized internationalization work, since internationalization expertise is not available in the respective protocols or working groups. Accordingly, the IESG formed the PRECIS working group to undertake the task.

Notwithstanding the desire evident in [NEWPREP] and the chartering of a working group, IDNA2008 may be a poor model for what other protocols ought to do, because it is designed to support an old protocol that is designed to operate on the scale of the entire Internet. Moreover, IDNA2008 is intended to be deployed without any

change to the base DNS protocol. Other protocols may aim at deployment in more local environments, or may have protocol version negotiation built in.

5. Major Topics for Consideration

This section provides an overview of major topics that a Stringprep replacement needs to address. The headings correspond roughly with categories under which known Stringprep-using protocol RFCs have been evaluated. For the details of those evaluations, see Appendix A.

5.1. Comparison

5.1.1. Types of Identifiers

Following [ID-COMP], it is possible to organize identifiers into three classes in respect of how they may be compared with one another:

Absolute Identifiers: Identifiers that can be compared byte-by-byte for equality.

Definite Identifiers: Identifiers that have a well-defined comparison algorithm on which all parties agree.

Indefinite Identifiers: Identifiers that have no single comparison algorithm on which all parties agree.

Definite Identifiers include cases like the comparison of Unicode code points in different encodings: they do not match byte for byte but can all be converted to a single encoding which then does match byte for byte. Indefinite Identifiers are sometimes algorithmically comparable by well-specified subsets of parties. For more discussion of these categories, see [ID-COMP].

The section on treating the existing known cases, Appendix A, uses the categories above.

5.1.2. Effect of Comparison

The three classes of comparison style outlined in Section 5.1.1 may have different effects when applied. It is necessary to evaluate the effects if a comparison results in a false positive or a false negative, especially in terms of the consequences to security and usability.

5.2. Dealing with Characters

This section outlines a range of issues having to do with characters in the target protocols, the ways in which IDNA2008 might be a good analogy to other protocols, and ways in which it might be a poor one.

5.2.1. Case Folding, Case Sensitivity, and Case Preservation

In IDNA2003, labels are always mapped to lowercase before the Punycode transformation. In IDNA2008, there is no mapping at all: input is either a valid U-label or it is not. At the same time, uppercase characters are by definition not valid U-labels, because they fall into the Unstable category (category B) of [RFC5892].

If there are protocols that require case be preserved, then the analogy with IDNA2008 will break down. Accordingly, existing protocols are to be evaluated according to the following criteria:

1. Does the protocol use case folding? For all blocks of code points or just for certain subsets?

2. Is the system or protocol case-sensitive?

3. Does the system or protocol preserve case?

5.2.2. Stringprep and NFKC

Stringprep profiles may use normalization. If they do, they use NFKC [UAX15] (most profiles do). It is not clear that NFKC is the right normalization to use in all cases. In [UAX15], there is the following observation regarding Normalization Forms KC and KD: "It is best to think of these Normalization Forms as being like uppercase or lowercase mappings: useful in certain contexts for identifying core meanings, but also performing modifications to the text that may not always be appropriate." In general, it can be said that NFKC is more aggressive about finding matches between code points than NFC. For things like the spelling of users' names, NFKC may not be the best form to use. At the same time, one of the nice things about NFKC is that it deals with the width of characters that are otherwise similar, by canonicalizing half-width to full-width. This mapping step can be crucial in practice. A replacement for Stringprep depends on analyzing the different use profiles and considering whether NFKC or NFC is a better normalization for each profile.

For the purposes of evaluating an existing example of Stringprep use, it is helpful to know whether it uses no normalization, NFKC, or NFC.

5.2.3. Character Mapping

Along with the case mapping issues raised in Section 5.2.1, there is the question of whether some characters are mapped either to other characters or to nothing during Stringprep. [RFC3454], Section 3, outlines a number of characters that are mapped to nothing, and also permits Stringprep profiles to define their own mappings.

5.2.4. Prohibited Characters

Along with case folding and other character mappings, many protocols have characters that are simply disallowed. For example, control characters and special characters such as "@" or "/" may be prohibited in a protocol.

One of the primary changes of IDNA2008 is in the way it approaches Unicode code points, using the new inclusion-based approach (see Section 1).

Because of the default assumption in IDNA2008 that a code point is not allowed by the protocol, it has more than one class of "allowed by the protocol"; this is unlike IDNA2003. While some code points are disallowed outright, some are allowed only in certain contexts. The reasons for the context-dependent rules have to do with the way some characters are used. For instance, the ZERO WIDTH JOINER and ZERO WIDTH NON-JOINER (ZWJ, U+200D and ZWNJ, U+200C) are allowed with contextual rules because they are required in some circumstances, yet are considered punctuation by Unicode and would therefore be DISALLOWED under the usual IDNA2008 derivation rules. The goal of IDNA2008 is to provide the widest repertoire of code points possible and consistent with the traditional DNS "LDH" (letters, digits, hyphen) rule (see [RFC0952]), trusting to the operators of individual zones to make sensible (and usually more restrictive) policies for their zones.

5.2.5. Internal Structure, Delimiters, and Special Characters

IDNA2008 has a special problem with delimiters, because the delimiter "character" in the DNS wire format is not really part of the data. In DNS, labels are not separated exactly; instead, a label carries with it an indicator that says how long the label is. When the label is displayed in presentation format as part of a fully qualified domain name, the label separator FULL STOP, U+002E (.) is used to break up the labels. But because that label separator does not travel with the wire format of the domain name, there is no way to encode a different, "internationalized" separator in IDNA2008.

Other protocols may include characters with similar special meaning within the protocol. Common characters for these purposes include FULL STOP, U+002E (.); COMMERCIAL AT, U+0040 (@); HYPHEN-MINUS, U+002D (-); SOLIDUS, U+002F (/); and LOW LINE, U+005F (_). The mere inclusion of such a character in the protocol is not enough for it to be considered similar to another protocol using the same character; instead, handling of the character must be taken into consideration as well.

An important issue to tackle here is whether it is valuable to map to or from these special characters as part of the Stringprep replacement. In some locales, the analogue to FULL STOP, U+002E is some other character, and users may expect to be able to substitute their normal stop for FULL STOP, U+002E. At the same time, there are predictability arguments in favor of treating identifiers with FULL STOP, U+002E in them just the way they are treated under IDNA2008.

5.2.6. Restrictions Because of Glyph Similarity

Homoglyphs are similarly (or identically) rendered glyphs of different code points. For DNS names, homoglyphs may enable phishing. If a protocol requires some visual comparison by end-users, then the issue of homoglyphs is to be considered. In the DNS context, these issues are documented in [RFC5894] and [RFC4690]. However, IDNA2008 does not have a mechanism to deal with them, trusting DNS zone operators to enact sensible policies for the subset of Unicode they wish to support, given their user community. A similar policy/protocol split may not be desirable in every protocol.

5.3. Where the Data Comes from and Where It Goes

5.3.1. User Input and the Source of Protocol Elements

Some protocol elements are provided by users, and others are not. Those that are not may presumably be subject to greater restrictions, whereas those that users provide likely need to permit the broadest range of code points. The following questions are helpful:

1. Do users input the strings directly?

2. If so, how? (keyboard, stylus, voice, copy-paste, etc.)

3. Where do we place the dividing line between user interface and protocol? (see [RFC5895])

5.3.2. User Output

Just as only some protocol elements are expected to be entered directly by users, only some protocol elements are intended to be consumed directly by users. It is important to know how users are expected to be able to consume the protocol elements, because different environments present different challenges. An element that is only ever delivered as part of a vCard remains in machine-readable format, so the problem of visual confusion is not a great one. Is the protocol element published as part of a vCard, a web directory, on a business card, or on "the side of a bus"? Do users use the protocol element as an identifier (which means that they might enter it again in some other context)? (See also Section 5.2.6.)

5.3.3. Operations

Some strings are useful as part of the protocol but are not used as input to other operations (for instance, purely informative or descriptive text). Other strings are used directly as input to other operations (such as cryptographic hash functions), or are used together with other strings to (such as concatenating a string with some others to form a unique identifier).

5.3.3.1. String Classes

Strings often have a similar function in different protocols. For instance, many different protocols contain user identifiers or passwords. A single profile for all such uses might be desirable.

Often, a string in a protocol is effectively a protocol element from another protocol. For instance, different systems might use the same credentials database for authentication.

5.3.3.2. Community Considerations

A Stringprep replacement that does anything more than just update Stringprep to the latest version of Unicode will probably entail some changes. It is important to identify the willingness of the protocol-using community to accept backwards-incompatible changes. By the same token, it is important to evaluate the desire of the community for features not available under Stringprep.

5.3.3.3. Unicode Incompatible Changes

IDNA2008 uses an algorithm to derive the validity of a Unicode code point for use under IDNA2008. It does this by using the properties of each code point to test its validity.

This approach depends crucially on the idea that code points, once valid for a protocol profile, will not later be made invalid. That is not a guarantee currently provided by Unicode. Properties of code points may change between versions of Unicode. Rarely, such a change could cause a given code point to become invalid under a protocol profile, even though the code point would be valid with an earlier version of Unicode. This is not merely a theoretical possibility, because it has occurred [RFC6452].

Accordingly, as in IDNA2008, a Stringprep replacement that intends to be Unicode version agnostic will need to work out a mechanism to address cases where incompatible changes occur because of new Unicode versions.

6. Considerations for Stringprep Replacement

The above suggests the following guidance:

- A Stringprep replacement should be defined.
- The replacement should take an approach similar to IDNA2008 (e.g., by using properties of code points instead of whitelisting of code points), in that it enables better Unicode agility.
- Protocols share similar characteristics of strings. Therefore, defining internationalization preparation algorithms for the smallest set of string classes may be sufficient for most cases, providing coherence among a set of related protocols or protocols where identifiers are exchanged.
- The sets of string classes need to be evaluated according to the considerations that make up the headings in Section 5
- It is reasonable to limit scope to Unicode code points and rule the mapping of data from other character encodings outside the scope of this effort.
- The replacement ought to at least provide guidance to applications using the replacement on how to handle protocol incompatibilities resulting from changes to Unicode. In an ideal world, the Stringprep replacement would handle the changes automatically, but it appears that such automatic handling would require magic and cannot be expected.
- Compatibility within each protocol between a technique that is Stringprep-based and the technique’s replacement has to be considered very carefully.

Existing deployments already depend on Stringprep profiles. Therefore, a replacement must consider the effects of any new strategy on existing deployments. By way of comparison, it is worth noting that some characters were acceptable in IDNA labels under IDNA2003, but are not protocol-valid under IDNA2008 (and conversely); disagreement about what to do during the transition has resulted in different approaches to mapping. Different implementers may make different decisions about what to do in such cases; this could have interoperability effects. It is necessary to trade better support for different linguistic environments against the potential side effects of backward incompatibility.

7. Security Considerations

This document merely states what problems are to be solved and does not define a protocol. There are undoubtedly security implications of the particular results that will come from the work to be completed. Moreover, the Stringprep Security Considerations [RFC3454] Section applies. See also the analysis in the subsections of Appendix B, below.

8. Acknowledgements

This document is the product of the PRECIS IETF Working Group, and participants in that working group were helpful in addressing issues with the text.

Specific contributions came from David Black, Alan DeKok, Simon Josefsson, Bill McQuillan, Alexey Melnikov, Peter Saint-Andre, Dave Thaler, and Yoshiro Yoneya.

Dave Thaler provided the "buckets" insight in Section 5.1.1, central to the organization of the problem.

Evaluations of Stringprep profiles that are included in Appendix B were done by David Black, Alexey Melnikov, Peter Saint-Andre, and Dave Thaler.

9. Informative References

[78PRECIS] Blanchet, M., "PRECIS Framework", Proceedings of IETF 78, July 2010, <http://www.ietf.org/proceedings/78/slides/precis-2.pdf>.

[ID-COMP] Thaler, D., Ed., "Issues in Identifier Comparison for Security Purposes", Work in Progress, March 2013.

[NEWPREP] "Newprep BoF Meeting Minutes", March 2010, <http://www.ietf.org/proceedings/77/minutes/newprep.txt>.

[RFC0952] Harrenstien, K., Stahl, M., and E. Feinler, "DoD Internet host table specification", RFC 952, October 1985.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC3454] Hoffman, P. and M. Blanchet, "Preparation of Internationalized Strings ("stringprep")", RFC 3454, December 2002.

[RFC3490] Faltstrom, P., Hoffman, P., and A. Costello, "Internationalizing Domain Names in Applications (IDNA)", RFC 3490, March 2003.

[RFC3491] Hoffman, P. and M. Blanchet, "Nameprep: A Stringprep Profile for Internationalized Domain Names (IDN)", RFC 3491, March 2003.

[RFC3492] Costello, A., "Punycode: A Bootstring encoding of Unicode for Internationalized Domain Names in Applications (IDNA)", RFC 3492, March 2003.

[RFC3530] Shepler, S., Callaghan, B., Robinson, D., Thurlow, R., Beame, C., Eisler, M., and D. Noveck, "Network File System (NFS) version 4 Protocol", RFC 3530, April 2003.

[RFC3722] Bakke, M., "String Profile for Internet Small Computer Systems Interface (iSCSI) Names", RFC 3722, April 2004.

[RFC3748] Aboba, B., Blunk, L., Vollbrecht, J., Carlson, J., and H. Levkowetz, "Extensible Authentication Protocol (EAP)", RFC 3748, June 2004.

[RFC3920] Saint-Andre, P., Ed., "Extensible Messaging and Presence Protocol (XMPP): Core", RFC 3920, October 2004.

[RFC3922] Saint-Andre, P., "Mapping the Extensible Messaging and Presence Protocol (XMPP) to Common Presence and Instant Messaging (CPIM)", RFC 3922, October 2004.

[RFC4011] Waldbusser, S., Saperia, J., and T. Hongal, "Policy Based Management MIB", RFC 4011, March 2005.

[RFC4013] Zeilenga, K., "SASLprep: Stringprep Profile for User Names and Passwords", RFC 4013, February 2005.

[RFC4279] Eronen, P. and H. Tschofenig, "Pre-Shared Key Ciphersuites for Transport Layer Security (TLS)", RFC 4279, December 2005.

[RFC4314] Melnikov, A., "IMAP4 Access Control List (ACL) Extension", RFC 4314, December 2005.

[RFC4422] Melnikov, A. and K. Zeilenga, "Simple Authentication and Security Layer (SASL)", RFC 4422, June 2006.

[RFC4505] Zeilenga, K., "Anonymous Simple Authentication and Security Layer (SASL) Mechanism", RFC 4505, June 2006.

[RFC4511] Sermersheim, J., "Lightweight Directory Access Protocol (LDAP): The Protocol", RFC 4511, June 2006.

[RFC4513] Harrison, R., "Lightweight Directory Access Protocol (LDAP): Authentication Methods and Security Mechanisms", RFC 4513, June 2006.

[RFC4518] Zeilenga, K., "Lightweight Directory Access Protocol (LDAP): Internationalized String Preparation", RFC 4518, June 2006.

[RFC4616] Zeilenga, K., "The PLAIN Simple Authentication and Security Layer (SASL) Mechanism", RFC 4616, August 2006.

[RFC4643] Vinocur, J. and K. Murchison, "Network News Transfer Protocol (NNTP) Extension for Authentication", RFC 4643, October 2006.

[RFC4683] Park, J., Lee, J., Lee, H., Park, S., and T. Polk, "Internet X.509 Public Key Infrastructure Subject Identification Method (SIM)", RFC 4683, October 2006.

[RFC4690] Klensin, J., Faltstrom, P., Karp, C., and IAB, "Review and Recommendations for Internationalized Domain Names (IDNs)", RFC 4690, September 2006.

[RFC4790] Newman, C., Duerst, M., and A. Gulbrandsen, "Internet Application Protocol Collation Registry", RFC 4790, March 2007.

[RFC4954] Siemborski, R. and A. Melnikov, "SMTP Service Extension for Authentication", RFC 4954, July 2007.

[RFC5034] Siemborski, R. and A. Menon-Sen, "The Post Office Protocol (POP3) Simple Authentication and Security Layer (SASL) Authentication Mechanism", RFC 5034, July 2007.

[RFC5051] Crispin, M., "i;unicode-casemap - Simple Unicode Collation Algorithm", RFC 5051, October 2007.

[RFC5054] Taylor, D., Wu, T., Mavrogiannopoulos, N., and T. Perrin, "Using the Secure Remote Password (SRP) Protocol for TLS Authentication", RFC 5054, November 2007.

[RFC5122] Saint-Andre, P., "Internationalized Resource Identifiers (IRIs) and Uniform Resource Identifiers (URIs) for the Extensible Messaging and Presence Protocol (XMPP)", RFC 5122, February 2008.

[RFC5280] Cooper, D., Santesson, S., Farrell, S., Boeyen, S., Housley, R., and W. Polk, "Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile", RFC 5280, May 2008.

[RFC5456] Spencer, M., Capouch, B., Guy, E., Miller, F., and K. Shumard, "IAX: Inter-Asterisk eXchange Version 2", RFC 5456, February 2010.

[RFC5661] Shepler, S., Eisler, M., and D. Noveck, "Network File System (NFS) Version 4 Minor Version 1 Protocol", RFC 5661, January 2010.

[RFC5802] Newman, C., Menon-Sen, A., Melnikov, A., and N. Williams, "Salted Challenge Response Authentication Mechanism (SCRAM) SASL and GSS-API Mechanisms", RFC 5802, July 2010.

[RFC5804] Melnikov, A. and T. Martin, "A Protocol for Remotely Managing Sieve Scripts", RFC 5804, July 2010.

[RFC5890] Klensin, J., "Internationalized Domain Names for Applications (IDNA): Definitions and Document Framework", RFC 5890, August 2010.

[RFC5891] Klensin, J., "Internationalized Domain Names in Applications (IDNA): Protocol", RFC 5891, August 2010.

[RFC5892] Faltstrom, P., "The Unicode Code Points and Internationalized Domain Names for Applications (IDNA)", RFC 5892, August 2010.

[RFC5893] Alvestrand, H. and C. Karp, "Right-to-Left Scripts for Internationalized Domain Names for Applications (IDNA)", RFC 5893, August 2010.

[RFC5894] Klensin, J., "Internationalized Domain Names for Applications (IDNA): Background, Explanation, and Rationale", RFC 5894, August 2010.

[RFC5895] Resnick, P. and P. Hoffman, "Mapping Characters for Internationalized Domain Names in Applications (IDNA) 2008", RFC 5895, September 2010.

[RFC6120] Saint-Andre, P., "Extensible Messaging and Presence Protocol (XMPP): Core", RFC 6120, March 2011.

[RFC6365] Hoffman, P. and J. Klensin, "Terminology Used in Internationalization in the IETF", BCP 166, RFC 6365, September 2011.

[RFC6452] Faltstrom, P. and P. Hoffman, "The Unicode Code Points and Internationalized Domain Names for Applications (IDNA) - Unicode 6.0", RFC 6452, November 2011.

[UAX15] "Unicode Standard Annex #15: Unicode Normalization Forms", UAX 15, September 2009.

[Unicode61] The Unicode Consortium. The Unicode Standard, Version 6.1.0, (Mountain View, CA: The Unicode Consortium, 2012. ISBN 978-1-936213-02-3). <http://www.unicode.org/versions/Unicode6.1.0/>.

Appendix A. Classification of Stringprep Profiles

A number of the known cases of Stringprep use were evaluated during the preparation of this document. The known cases are here described in two ways. The types of identifiers the protocol uses is first called out in the ID type column (from Section 5.1.1) using the short forms "a" for Absolute, "d" for Definite, and "i" for Indefinite. Next, there is a column that contains an "i" if the protocol string comes from user input, an "o" if the protocol string becomes user-facing output, "b" if both are true, and "n" if neither is true.

RFC	IDtype	User?
3722	a	b
3748	-	-
3920	a,d	b
4505	a	i
4314	a,d	b
4954	a,d	b
5034	a,d	b
5804	a,d	b

Table 1

Appendix B. Evaluation of Stringprep Profiles

This section is a summary of evaluation of Stringprep profiles that was done to get a good understanding of the usage of Stringprep. This summary is by no means normative nor the actual evaluations themselves. A template was used for reviewers to get a coherent view of all evaluations.

B.1. iSCSI Stringprep Profile: RFC 3720, RFC 3721, RFC 3722

Description: An iSCSI session consists of an initiator (i.e., host or server that uses storage) communicating with a target (i.e., a storage array or other system that provides storage). Both the iSCSI initiator and target are named by iSCSI names. The iSCSI Stringprep profile is used for iSCSI names.

How it is used: iSCSI initiators and targets (see above). They can also be used to identify SCSI ports (these are software entities in the iSCSI protocol, not hardware ports) and iSCSI logical units (storage volumes), although both are unusual in practice.

What entities create these identifiers? Generally, a human user (1) configures an automated system (2) that generates the names. Advance configuration of the system is required due to the embedded use of external unique identifier (from the DNS or IEEE).

How is the string input in the system? Keyboard and copy-paste are common. Copy-paste is common because iSCSI names are long enough to be problematic for humans to remember, causing use of email, sneaker-net, text files, etc., to avoid mistype mistakes.

Where do we place the dividing line between user interface and protocol? The iSCSI protocol requires that all internationalization string preparation occur in the user interface. The iSCSI protocol treats iSCSI names as opaque identifiers that are compared byte-by-byte for equality. iSCSI names are generally not checked for correct formatting by the protocol.

What entities enforce the rules? There are no iSCSI-specific enforcement entities, although the use of unique identifier information in the names relies on DNS registrars and the IEEE Registration Authority.

Comparison: Byte-by-byte.

Case Folding, Sensitivity, Preservation: Case folding is required for the code blocks specified in RFC 3454, Table B.2. The overall iSCSI naming system (UI + protocol) is case-insensitive.

What is the impact if the comparison results in a false positive? Potential access to the wrong storage.

- If the initiator has no access to the wrong storage, an authentication failure is the probable result.

- If the initiator has access to the wrong storage, the resulting misidentification could result in use of the wrong data and possible corruption of stored data.

What is the impact if the comparison results in a false negative? Denial of authorized storage access.

What are the security impacts? iSCSI names may be used as the authentication identities for storage systems. Comparison problems could result in authentication problems, although note that authentication failure ameliorates some of the false positive cases.

Normalization: NFKC, as specified by RFC 3454.

Mapping: Yes, as specified by Table B.1 in RFC 3454.

Disallowed Characters: Only the following characters are allowed:
- ASCII dash, dot, colon
- ASCII lowercase letters and digits
- Unicode lowercase characters as specified by RFC 3454.

All other characters are disallowed.

Which other strings or identifiers are these most similar to?
None -- iSCSI names are unique to iSCSI.

Are these strings or identifiers sometimes the same as strings or identifiers from other protocols? No.

Does the identifier have internal structure that needs to be respected? Yes. ASCII dot, dash, and colon are used for internal name structure. These are not reserved characters, in that they can occur in the name in locations other than those used for structuring purposes (e.g., only the first occurrence of a colon character is structural, others are not).

How are users exposed to these strings? How are they published?
iSCSI names appear in server and storage system configuration interfaces. They also appear in system logs.

Is the string / identifier used as input to other operations?
Effectively, no. The rarely used port and logical unit names involve concatenation, which effectively extends a unique iSCSI name for a target to uniquely identify something within that target.

How much tolerance for change from existing Stringprep approach?
Good tolerance; the community would prefer that internationalization experts solve internationalization problems.

How strong a desire for change (e.g., for Unicode agility)? Unicode agility is desired, in principle, as long as nothing significant breaks.

B.2. SMTP/POP3/ManageSieve Stringprep Profiles: RFC 4954, RFC 5034, RFC 5804

Description: Authorization identity (user identifier) exchanged during SASL authentication: AUTH (SMTP/POP3) or AUTHENTICATE (ManageSieve) command.

How It's Used: Used for proxy authorization, e.g., to [lawfully] impersonate a particular user after a privileged authentication.

Who Generates It:
- Typically generated by email system administrators using some tools/conventions, sometimes from some backend database.
- In some setups, human users can register their own usernames (e.g., webmail self-registration).

User Input Methods:
- typing or selecting from a list
- copy and paste
- voice input
- in configuration files or on the command line

Enforcement: Rules enforced by server / add-on service (e.g., gateway service) on registration of account.

Comparison Method: "Type 1" (byte-for-byte) or "Type 2" (compare by a common algorithm that everyone agrees on (e.g., normalize and then compare the result byte-by-byte).

Case Folding, Sensitivity, Preservation: Most likely case-sensitive. Exact requirements on case-sensitivity/case-preservation depend on a specific implementation, e.g., an implementation might treat all user identifiers as case-insensitive (or case-insensitive for US-ASCII subset only).

Impact of Comparison: False positives: an unauthorized user is allowed email service access (login). False negatives: an authorized user is denied email service access.

Normalization: NFKC (as per RFC 4013).

Mapping: (see Section 2 of RFC 4013 for the full list) Non-ASCII spaces are mapped to space, etc.

Disallowed Characters: (see Section 2 of RFC 4013 for the full list) Unicode Control characters, etc.

String Classes: Simple username. See Section 2 of RFC 4013 for details on restrictions. Note that some implementations allow spaces in these. While implementations are not required to use a specific format, an authorization identity frequently has the same format as an email address (and Email Address Internationalization (EAI) email address in the future), or as a left hand side of an email address. Note: whatever is recommended for SMTP/POP/

ManageSieve authorization identity should also be used for IMAP authorization identities, as IMAP/POP3/SMTP/ManageSieve are frequently implemented together.

Internal Structure: None

User Output: Unlikely, but possible. For example, if it is the same as an email address.

Operations: Sometimes concatenated with other data and then used as input to a cryptographic hash function.

How much tolerance for change from existing Stringprep approach? Not sure.

Background Information:
In RFC 5034, when describing the POP3 AUTH command:

> The authorization identity generated by the SASL exchange is a simple username, and SHOULD use the SASLprep profile (see [RFC4013]) of the StringPrep algorithm (see [RFC3454]) to prepare these names for matching. If preparation of the authorization identity fails or results in an empty string (unless it was transmitted as the empty string), the server MUST fail the authentication.

In RFC 4954, when describing the SMTP AUTH command:

> The authorization identity generated by this [SASL] exchange is a "simple username" (in the sense defined in [SASLprep]), and both client and server SHOULD (*) use the [SASLprep] profile of the [StringPrep] algorithm to prepare these names for transmission or comparison. If preparation of the authorization identity fails or results in an empty string (unless it was transmitted as the empty string), the server MUST fail the authentication.
>
> (*) Note: Future revision of this specification may change this requirement to MUST. Currently, the SHOULD is used in order to avoid breaking the majority of existing implementations.

In RFC 5804, when describing the ManageSieve AUTHENTICATE command:

> The authorization identity generated by this [SASL] exchange is a "simple username" (in the sense defined in [SASLprep]), and both client and server MUST use the [SASLprep] profile of the [StringPrep] algorithm to prepare these names for transmission or comparison. If preparation of the authorization identity fails or results in an empty string (unless it was transmitted as the empty string), the server MUST fail the authentication.

B.3. IMAP Stringprep Profiles for Usernames: RFC 4314, RFC 5738

Evaluation Note: These documents have 2 types of strings (usernames and passwords), so there are two separate templates.

Description: "username" parameter to the IMAP LOGIN command, identifiers in IMAP Access Control List (ACL) commands. Note that any valid username is also an IMAP ACL identifier, but IMAP ACL identifiers can include other things like the name of a group of users.

How It's Used: Used for authentication (Usernames), or in IMAP Access Control Lists (Usernames or Group names).

Who Generates It:
- Typically generated by email system administrators using some tools/conventions, sometimes from some backend database.
- In some setups, human users can register own usernames (e.g., webmail self-registration).

User Input Methods:
- typing or selecting from a list
- copy and paste
- voice input
- in configuration files or on the command line

Enforcement: Rules enforced by server / add-on service (e.g., gateway service) on registration of account.

Comparison Method: "Type 1" (byte-for-byte) or "Type 2" (compare by a common algorithm that everyone agrees on (e.g., normalize and then compare the result byte-by-byte).

Case Folding, Sensitivity, Preservation: Most likely case-sensitive. Exact requirements on case-sensitivity/case-preservation depend on a specific implementation, e.g., an implementation might treat all user identifiers as case-insensitive (or case-insensitive for US-ASCII subset only).

Impact of Comparison: False positives: an unauthorized user is allowed IMAP access (login), privileges improperly granted (e.g., access to a specific mailbox, ability to manage ACLs for a mailbox). False negatives: an authorized user is denied IMAP access, unable to use granted privileges (e.g., access to a specific mailbox, ability to manage ACLs for a mailbox).

Normalization: NFKC (as per RFC 4013)

Mapping: (see Section 2 of RFC 4013 for the full list) Non-ASCII spaces are mapped to space.

Disallowed Characters: (see Section 2 of RFC 4013 for the full list) Unicode Control characters, etc.

String Classes: Simple username. See Section 2 of RFC 4013 for details on restrictions. Note that some implementations allow spaces in these. While IMAP implementations are not required to use a specific format, an IMAP username frequently has the same format as an email address (and EAI email address in the future), or as a left hand side of an email address. Note: whatever is recommended for the IMAP username should also be used for ManageSieve, POP3 and SMTP authorization identities, as IMAP/POP3/SMTP/ManageSieve are frequently implemented together.

Internal Structure: None.

User Output: Unlikely, but possible. For example, if it is the same as an email address, access control lists (e.g. in IMAP ACL extension), both when managing membership and listing membership of existing access control lists. Often shows up as mailbox names (under Other Users IMAP namespace).

Operations: Sometimes concatenated with other data and then used as input to a cryptographic hash function.

How much tolerance for change from existing Stringprep approach? Not sure. Non-ASCII IMAP usernames are currently prohibited by IMAP (RFC 3501). However, they are allowed when used in IMAP ACL extension.

B.4. IMAP Stringprep Profiles for Passwords: RFC 5738

Description: "Password" parameter to the IMAP LOGIN command.

How It's Used: Used for authentication (Passwords).

Who Generates It: Either generated by email system administrators using some tools/conventions, or specified by the human user.

User Input Methods:
- typing or selecting from a list
- copy and paste
- voice input
- in configuration files or on the command line

Enforcement: Rules enforced by server / add-on service (e.g., gateway service or backend database) on registration of account.

Comparison Method: "Type 1" (byte-for-byte).

Case Folding, Sensitivity, Preservation: Most likely case-sensitive.

Impact of Comparison: False positives: an unauthorized user is allowed IMAP access (login). False negatives: an authorized user is denied IMAP access.

Normalization: NFKC (as per RFC 4013).

Mapping: (see Section 2 of RFC 4013 for the full list) Non-ASCII spaces are mapped to space.

Disallowed Characters: (see Section 2 of RFC 4013 for the full list) Unicode Control characters, etc.

String Classes: Currently defined as "simple username" (see Section 2 of RFC 4013 for details on restrictions); however, this is likely to be a different class from usernames. Note that some implementations allow spaces in these. Password in all email related protocols should be treated in the same way. Same passwords are frequently shared with web, IM, and etc. applications.

Internal Structure: None.

User Output: Text of email messages (e.g. in "you forgot your password" email messages), web page / directory, side of the bus / in ads -- possible.

Operations: Sometimes concatenated with other data and then used as input to a cryptographic hash function. Frequently stored as is, or hashed.

How much tolerance for change from existing Stringprep approach? Not sure. Non-ASCII IMAP passwords are currently prohibited by IMAP (RFC 3501); however, they are likely to be in widespread use.

Background Information:
RFC 5738, Section 5 ("UTF8=USER Capability"):

> If the "UTF8=USER" capability is advertised, that indicates the server accepts UTF-8 user names and passwords and applies SASLprep [RFC4013] to both arguments of the LOGIN command. The server MUST reject UTF-8 that fails to comply with the formal syntax in RFC 3629 [RFC3629] or if it encounters Unicode characters listed in Section 2.3 of SASLprep RFC 4013 [RFC4013].

RFC 4314, Section 3 ("Access control management commands and responses"):

> Servers, when processing a command that has an identifier as a parameter (i.e., any of SETACL, DELETEACL, and LISTRIGHTS commands), SHOULD first prepare the received identifier using "SASLprep" profile [SASLprep] of the "stringprep" algorithm [Stringprep]. If the preparation of the identifier fails or results in an empty string, the server MUST refuse to perform the command with a BAD response. Note that Section 6 recommends additional identifier’s verification steps.

RFC 4314, Section 6 ("Security Considerations"):

> This document relies on [SASLprep] to describe steps required to perform identifier canonicalization (preparation). The preparation algorithm in SASLprep was specifically designed such that its output is canonical, and it is well-formed. However, due to an anomaly [PR29] in the specification of Unicode normalization, canonical equivalence is not guaranteed for a select few character sequences. Identifiers prepared with SASLprep can be stored and returned by an ACL server. The anomaly affects ACL manipulation and evaluation of identifiers containing the selected character sequences. These sequences, however, do not appear in well-formed text. In order to address this problem, an ACL server MAY reject identifiers containing sequences described in [PR29] by sending the tagged

BAD response. This is in addition to the requirement to reject identifiers that fail SASLprep preparation as described in Section 3.

B.5. Anonymous SASL Stringprep Profiles: RFC 4505

Description: RFC 4505 defines a "trace" field:

Comparison: this field is not intended for comparison (only used for logging)

Case folding; case-sensitivity, preserve case: No case folding/ case-sensitive

Do users input the strings directly? Yes. Possibly entered in configuration UIs, or on a command line. Can also be stored in configuration files. The value can also be automatically generated by clients (e.g., a fixed string is used, or a user's email address).

How users input strings? Keyboard/voice, stylus (pick from a list). Copy-paste - possibly.

Normalization: None.

Disallowed Characters: Control characters are disallowed. (See Section 3 of RFC 4505).

Which other strings or identifiers are these most similar to? RFC 4505 says that the trace "should take one of two forms: an Internet email address, or an opaque string that does not contain the '@' (U+0040) character and that can be interpreted by the system administrator of the client's domain". In practice, this is a free-form text, so it belongs to a different class from "email address" or "username".

Are these strings or identifiers sometimes the same as strings or identifiers from other protocols (e.g., does an IM system sometimes use the same credentials database for authentication as an email system)? Yes: see above. However, there is no strong need to keep them consistent in the future.

How are users exposed to these strings, how are they published? No. However, the value can be seen in server logs.

Impacts of false positives and false negatives:
False positive: a user can be confused with another user.
False negative: two distinct users are treated as the same user.
But note that the trace field is not authenticated, so it can be easily falsified.

Tolerance of changes in the community: The community would be flexible.

Delimiters: No internal structure, but see comments above about frequent use of email addresses.

Background Information:
RFC 4505, Section 2 ("The Anonymous Mechanism"):

> The mechanism consists of a single message from the client to the server. The client may include in this message trace information in the form of a string of [UTF-8]-encoded [Unicode] characters prepared in accordance with [StringPrep] and the "trace" stringprep profile defined in Section 3 of this document. The trace information, which has no semantical value, should take one of two forms: an Internet email address, or an opaque string that does not contain the '@' (U+0040) character and that can be interpreted by the system administrator of the client's domain. For privacy reasons, an Internet email address or other information identifying the user should only be used with permission from the user.
>
> RFC 4505, Section 3 ('The "trace" Profile of "Stringprep"'):
> This section defines the "trace" profile of [StringPrep]. This profile is designed for use with the SASL ANONYMOUS Mechanism. Specifically, the client is to prepare the <message> production in accordance with this profile.
>
> The character repertoire of this profile is Unicode 3.2 [Unicode].
>
> No mapping is required by this profile.
>
> No Unicode normalization is required by this profile.
>
> The list of unassigned code points for this profile is that provided in Appendix A of [StringPrep]. Unassigned code points are not prohibited.

Characters from the following tables of [StringPrep] are prohibited:

- C.2.1 (ASCII control characters)
- C.2.2 (Non-ASCII control characters)
- C.3 (Private use characters)
- C.4 (Non-character code points)
- C.5 (Surrogate codes)
- C.6 (Inappropriate for plain text)
- C.8 (Change display properties are deprecated)
- C.9 (Tagging characters)

No additional characters are prohibited.

This profile requires bidirectional character checking per Section 6 of [StringPrep].

B.6. XMPP Stringprep Profiles for Nodeprep: RFC 3920

Description: Localpart of JabberID ("JID"), as in: localpart@domainpart/resourcepart

How It's Used:
- Usernames (e.g., stpeter@jabber.org)
- Chatroom names (e.g., precis@jabber.ietf.org)
- Publish-subscribe nodes
- Bot names

Who Generates It:
- Typically, end users via an XMPP client
- Sometimes created in an automated fashion

User Input Methods:
- typing
- copy and paste
- voice input
- clicking a URI/IRI

Enforcement: Rules enforced by server / add-on service (e.g., chatroom service) on registration of account, creation of room, etc.

Comparison Method: "Type 2" (common algorithm)

Case Folding, Sensitivity, Preservation:
- Strings are always folded to lowercase
- Case is not preserved

Impact of Comparison:
False positives:
- unable to authenticate at server (or authenticate to wrong account)
- add wrong person to buddy list
- join the wrong chatroom
- improperly grant privileges (e.g., chatroom admin)
- subscribe to wrong pubsub node
- interact with wrong bot
- allow communication with blocked entity

False negatives:
- unable to authenticate
- unable to add someone to buddy list
- unable to join desired chatroom
- unable to use granted privileges (e.g., chatroom admin)
- unable to subscribe to desired pubsub node
- unable to interact with desired bot
- disallow communication with unblocked entity

Normalization: NFKC

Mapping: Spaces are mapped to nothing

Disallowed Characters: ",&,',/,:,<,>,@

String Classes:
- Often similar to generic username
- Often similar to localpart of email address
- Sometimes same as localpart of email address

Internal Structure: None

User Output:
- vCard
- email signature
- web page / directory
- text of message (e.g., in a chatroom)

Operations: Sometimes concatenated with other data and then used as input to a cryptographic hash function

B.7. XMPP Stringprep Profiles for Resourceprep: RFC 3920

Description:
- Resourcepart of JabberID ("JID"), as in: localpart@domainpart/resourcepart
- Typically free-form text

How It's Used:
- Device / session names (e.g., stpeter@jabber.org/Home)
- Nicknames (e.g., precis@jabber.ietf.org/StPeter)

Who Generates It:
- Often human users via an XMPP client
- Often generated in an automated fashion by client or server

User Input Methods:
- typing
- copy and paste
- voice input
- clicking a URI/IRI

Enforcement: Rules enforced by server / add-on service (e.g., chatroom service) on account login, joining a chatroom, etc.

Comparison Method: "Type 2" (byte-for-byte)

Case Folding, Sensitivity, Preservation:
- Strings are never folded
- Case is preserved

Impact of Comparison:
False positives:
- interact with wrong device (e.g., for file transfer or voice call)
- interact with wrong chatroom participant
- improperly grant privileges (e.g., chatroom moderator)
- allow communication with blocked entity

False negatives:
- unable to choose desired chatroom nickname
- unable to use granted privileges (e.g., chatroom moderator)
- disallow communication with unblocked entity

Normalization: NFKC

Mapping: Spaces are mapped to nothing

Disallowed Characters: None

String Classes: Basically a free-form identifier

Internal Structure: None

User Output:
- text of message (e.g., in a chatroom)
- device names often not exposed to human users

Operations: Sometimes concatenated with other data and then used as input to a cryptographic hash function

B.8. EAP Stringprep Profiles: RFC 3748

Description: RFC 3748, Section 5, references Stringprep, but the WG did not agree with the text (was added by IESG) and there are no known implementations that use Stringprep. The main problem with that text is that the use of strings is a per-method concept, not a generic EAP concept and so RFC 3748 itself does not really use Stringprep, but individual EAP methods could. As such, the answers to the template questions are mostly not applicable, but a few answers are universal across methods. The list of IANA registered EAP methods is at <http://www.iana.org/assignments/eap-numbers/eap-numbers.xml>.

Comparison Methods: n/a (per-method)

Case Folding, Case-Sensitivity, Case Preservation: n/a (per-method)

Impact of comparison: A false positive results in unauthorized network access (and possibly theft of service if some else is billed). A false negative results in lack of authorized network access (no connectivity).

User input: n/a (per-method)

Normalization: n/a (per-method)

Mapping: n/a (per-method)

Disallowed characters: n/a (per-method)

String classes: Although some EAP methods may use a syntax similar to other types of identifiers, EAP mandates that the actual values must not be assumed to be identifiers usable with anything else.

Internal structure: n/a (per-method)

User output: Identifiers are never human displayed except perhaps as they’re typed by a human.

Operations: n/a (per-method)

Community considerations: There is no resistance to change for the base EAP protocol (as noted, the WG didn't want the existing text). However, actual use of Stringprep, if any, within specific EAP methods may have resistance. It is currently unknown whether any EAP methods use Stringprep.

Authors' Addresses

Marc Blanchet
Viagenie
246 Aberdeen
Quebec, QC G1R 2E1
Canada

EMail: Marc.Blanchet@viagenie.ca
URI: http://viagenie.ca

Andrew Sullivan
Dyn, Inc.
150 Dow St
Manchester, NH 03101
U.S.A.

EMail: asullivan@dyn.com