ICTビジネス戦略オンラインセミナー
「デジュール及びフォーラム標準に関する 国際標準化活動動向調査」

IETF が策定する国際化技術の標準化推進と国際化技術を活用するIoT 技術の動向調査

2021/01/14(木)

根本 貴弘

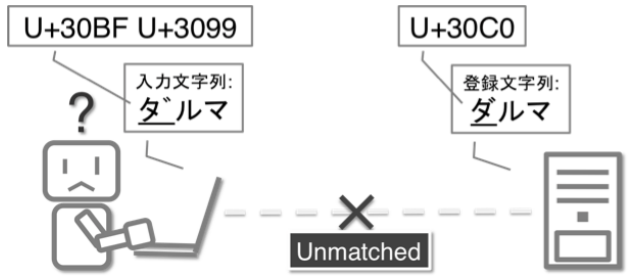
国立大学法人東京農工大学

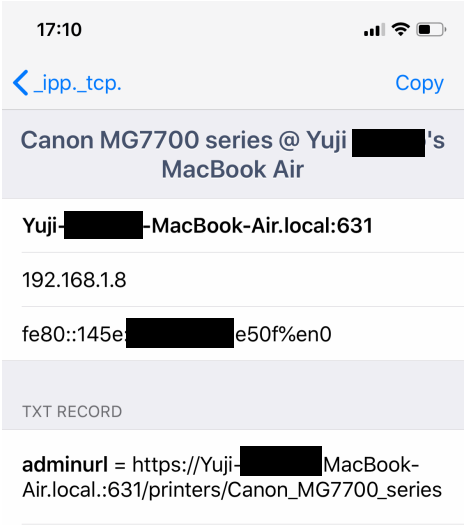
調査概要

- 調査機関
  - **Internet Engineering Task Force(IETF)**
- 調査テーマ
  - **IETF**が策定する国際化技術の標準化推進と国際化技術を活用する**IoT**技術の動向調査
- 調査概要
  1. **IETF**が策定する国際化技術の標準化推進として**IETF**の主要な国際化技術である**PRECIS Framework**の**Unicode 7.0**以降への対応提案
  2. **IoT**サービスの情報資源参照時における利便性や安全性，相互運用性の向上に必要となる国際化技術の観点からみた課題の共有

調査背景

- 使用者が任意に命名可能なモノを情報資源として扱う可能性のある，**IoT**サービスにおいて情報資源参照時や認証時における利便性及び安全性の向上のためには適切な国際化技術を使用する必要

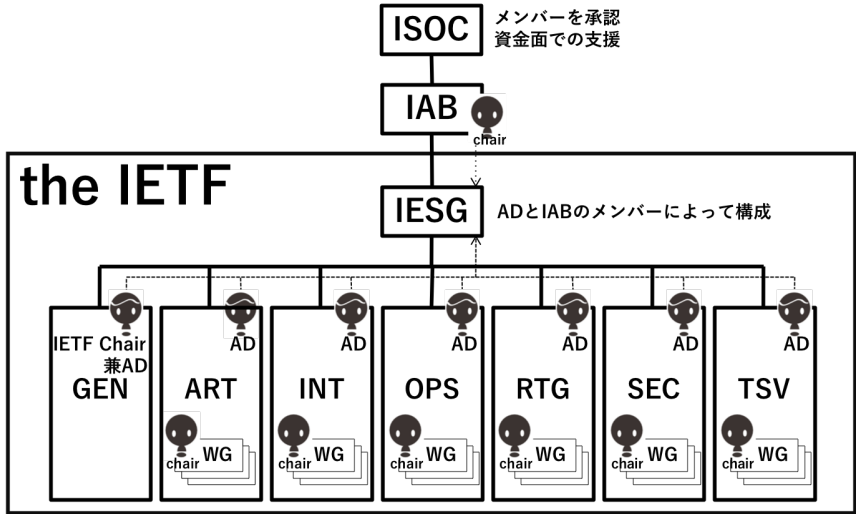




- **Unicode 7.0**以降への対応
  - **Unicode 10.0**で戸籍統一文字や住民基本台帳ネットワーク文字等の行政システムで利用される文字はほぼ一通り収録された文字も利用できない
- 情報システムにおける相互運用性を考慮し文字情報の整備が必要
  - **IoT**サービスに関するプロトコルで適切に国際化技術が検討されていない

IETF概要

- **Internet Engineering Task Force (IETF)**
  - インターネット技術に係る技術標準としての**Request for Comments (RFC)**の仕様策定のプロセスに責任を持つ団体
  - 技術的な内容及び作業に係る責任は**Internet Engineering Steering Group (IESG)**が担う
  - **Internet Architecture Board (IAB)**のタスクフォースの一つであり，**IAB**が定めるインターネットの標準化プロセスの方針に従い活動を行う
  - **IAB**の上位組織の**Internet Society (ISOC)**は資金面での支援
  - 長期的な資金調達計画等を行う**IETF LLC**を**2018**年**8**月に設立
- 代表的な**RFC**例
  - **IP (RFC791), TCP (RFC793), DNS (RFC1034, RFC1035)**．．．
- **IETF**の標準化過程に関する情報は全て公開
- **IETF**での標準化作業は，年**3**回開催される**IETF**会合中の**WG**会合や**WG**のメーリングリストでの議論を通じて行う



IETFにおける国際化の背景

- 国際化（**internationalization = i18n**）
- **RFC2130: The Report of the IAB Character Set Workshop held 29 February - 1 March, 1996**
  - インターネットは当初**ASCII**文字集合を前提に設計
  - インターネットの広がりが**1980**年代後半から加速，利用者範囲が拡大
  - **ASCII**文字集合以外の文字集合を利用したいという需要が増加
- **1990**年代初頭：**MIME**が標準化
  - 電子メールや**HTML**の本文で**ASCII**文字集合以外の文字が利用可能となった
- **1990**年代半ば：商用**ISP**とインターネット対応コンシューマ向け**OS**の登場により，インターネット利用者がさらに増加
  - 電子メールや**Web**によるコミュニケーションが日常的となる
- **1990**年代後半：インターネット上のサービスで使われる識別子（ドメイン名や電子メールアドレス）等で母国語で使用する文字を使用したいという需要が生まれる

IETFにおける国際化

- **RFC 6365: Terminology Used in Internationalization in the IETF (BCP 166)**
  - 著者**: P. Hoffman, J. Klensin**
  - 発行時期**: 2011年9月**
  - **Obsoletes: RFC 3536**
- **(**主にアプリケーション**)**プロトコルで非**ASCII**文字集合を扱えるようにすること
  - **“In the IETF, "internationalization" means to add or improve the handling of non-ASCII text in a protocol. ”**
  - **W3C**の**i18n**では，はじめから国際化を考慮して設計されたプロトコルが前提として考えられている
    - **“Internationalization is the design and development of a product, application or document content that enables easy localization for target audiences that vary in culture, region, or language.” [W3C-i18n-Def]**
- 様々な国や地域で使用される文字集合をプロトコルで利用可能とするための共通の枠組み

IETFにおける文字集合と言語

- **RFC 2277: IETF Policy on Character Sets and Languages (BCP 18)**
  - 著者**: H. Alvestrand**
  - 発行時期**: 1998**年**1**月
- **(**アプリケーション**)**プロトコルは「**UTF-8**」を扱えなくてはならない
  - **“Protocols MUST be able to use the UTF-8 charset, which consists ofthe ISO 10646 coded character set combined with the UTF-8 character encoding scheme, as defined in [10646] Annex R (published inAmendment 2), for all text.”**
- 「国際化における検討」が必要
  - **“In documents that deal with internationalization issues at all, a synopsis of the approaches chosen for internationalization SHOULD be collected into a section called "Internationalization considerations", and placed next to the Security Considerations section.”**

Unicode文字集合

- **Unicode**文字集合 **(ISO/IEC 10646)**
  - 各言語で使用する文字を全て収録することを目標に開発された文字集合
  - **1**文字を**21bit**で表現
    - **1,114,112**通りのビットパターンが表現可能
  - 現在でも改版作業が継続
  - 他の文字集合との互換性を考慮して設計

改版時期	バージョン番号	収録文字数
2002年 3月	3.2.0	95,221
:	:	:
2010年 10月	6.0.0	109,449
2012年 1月	6.1.0	110,181
2012年 9月	6.2.0	110,182
:	:	:
2020年 5月	13.0.0	143,859

- **IETF**では，**Unicode**文字集合は**UTF-8**として扱う
  - **RFC 3629: UTF-8, a transformation format of ISO 10646 (STD 63)**にて，**Unicode**文字集合**(ISO / IEC 10646)**を扱うためのエンコーディング方式“**UTF-8**”について説明

Unicodeを用いた国際化の検討事項

- 識別子等で**Unicode/UTF-8**を使用するを実現するために検討すべき課題
  - プロトコルで利用可能な文字の分類
  - 視覚的に紛らわしい文字の扱い
  - **bidi**文字列の扱い
  - **Unicode**の改版への追従方法

文字の方向	例
左から右（ラテン文字）	nemoto
右から左（アラビア文字）	نَيْمُوتو
右から左（ターナ文字）	ތެސްޓް
双方向（アラビア文字・ラテン文字）	تو**O**نَيْمُ
双方向（アラビア文字・数字）	يْمُو١٠

必要な変換処理	例	
文字種（大文字・小文字）	A (U+0041)	a (U+0061)
文字幅（全角・半角）	ｱ (U+FF71)	ア (U+30A2)
合成済文字・結合文字列	ガ (U+30AB U+3099)	ガ (U+30AC)
文脈依存文字	σ (U+03C3)	ς (U+03C2)
言語依存文字	İ (U+0130)	i (U+0069)
区切り文字	。 (U+3002)	. (U+03C2)
空白文字	(U+3000)	(U+0020)
見た目に見えない文字	SHY (U+00AD)	(Nothing)

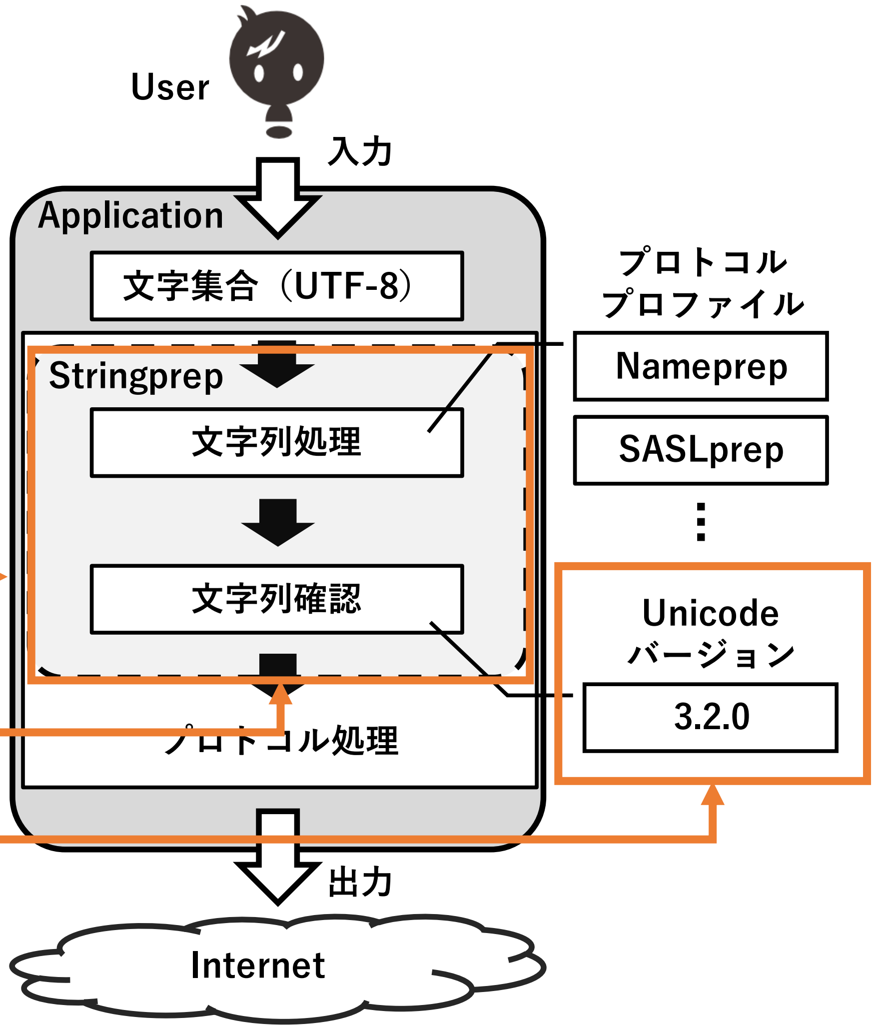
IETFが取り組んできた国際化技術における標準

- **Multipurpose Internet Mail Extensions (MIME)**
  - 電子メールや**HTML**の本文で**ASCII**文字集合以外の文字が扱うことが可能
  - **RFC2045**，**RFC2046**，**RFC2047**，**RFC2048(**現**RFC4288**，**RFC4289)**，**RFC 2049**
- **Internationalizing Domain Names in Applications (IDNA)**
  - 国際化ドメイン名（**IDNA2003**と**IDNA2008**がある）
  - **IDNA2003**：**RFC3490**，**RFC3491**，**RFC3492**
  - **IDNA2008**：**RFC5890**，**RFC5891**，**RFC5892**，**RFC5893**，**RFC5895**
- **Email Address Internationalization (EAI)**
  - 国際化電子メールアドレス
  - **RFC6530**，**RFC6531**，**RFC6532**，**RFC6533**，**RFC6855**，**RFC6856**，**RFC6857**，**RFC6858**
- **Stringprep**
  - 国際化文字列を扱うための枠組み
  - **RFC3454**
- **PRECIS Framework**
  - **Stringprep**に変わる国際化文字列を扱うための枠組み
  - **RFC8264**，**RFC8265**，**RFC8266**，**RFC6885**，**RFC7790**

Stringprep

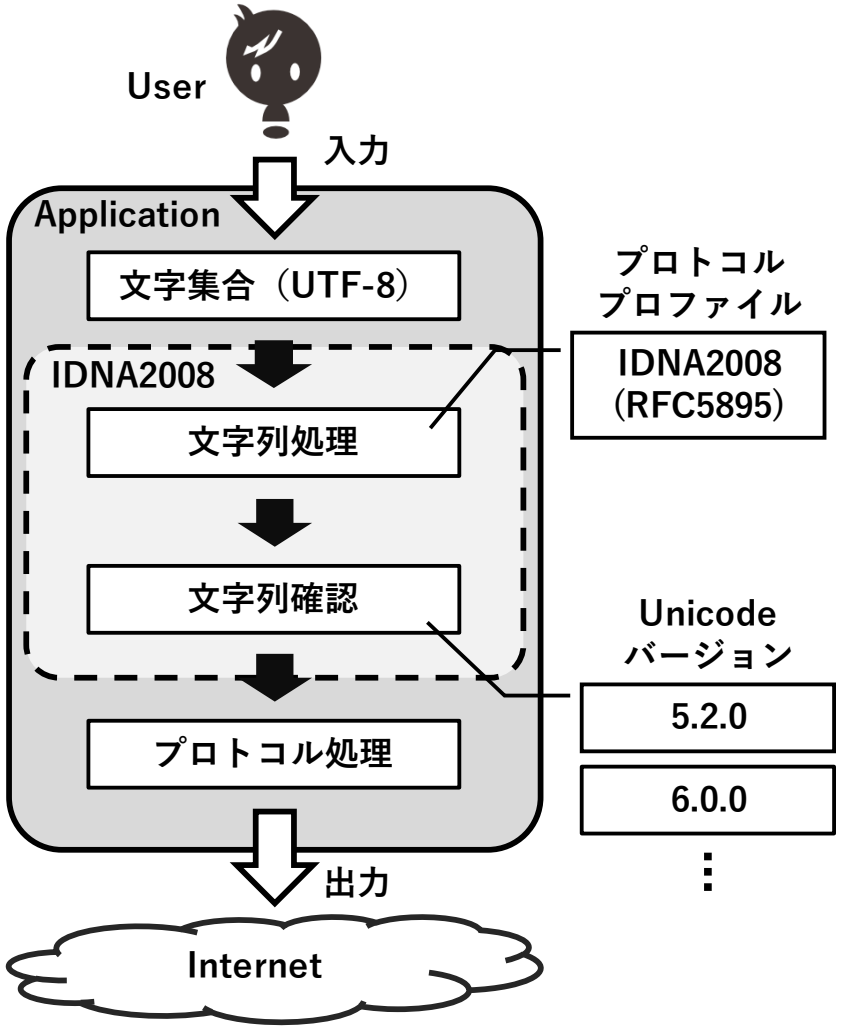
- 各通信プロトコルで国際化文字列を扱うための文字列処理の枠組み
  - 文字列変換処理と使用禁止文字を定義
  - 各プロトコルは必要な処理を選択して利用
- 課題
  - 文字列変換処理の問題**(Unicode**正規化形式**KC)**
  - 文字列確認方法の問題**(**双方向性，禁止文字**)**
  - <u>特定の文字集合のバージョンに依存</u>

    (㍾，㍽，㍼，㍻，~~㋿**(**新元号~~**))**



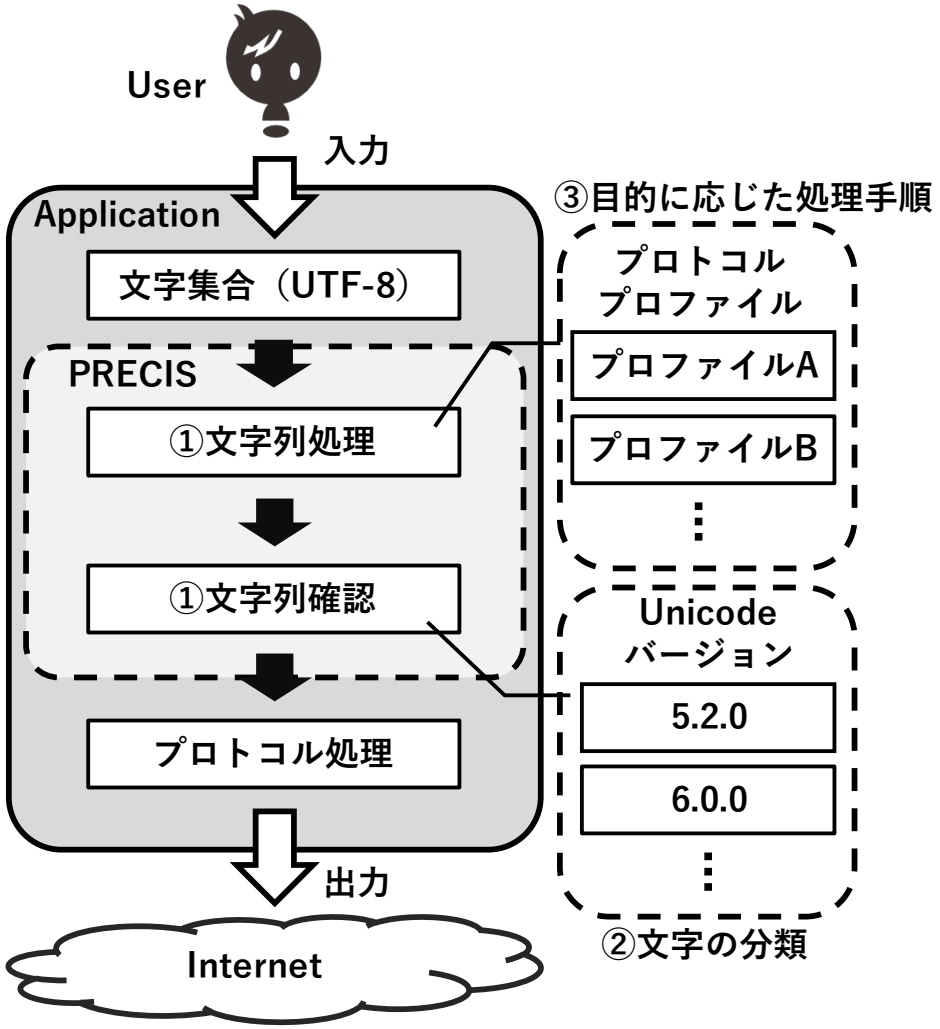
IDNA2008

- **Stringprep**を利用していた**IDNA2003**の課題を解決するために策定された，ドメイン名で国際化文字列を扱うための文字列処理の枠組み
  - 文字列変換処理と使用許可文字分類方法を定義
  - 各**Unicode**の改版に適応可能



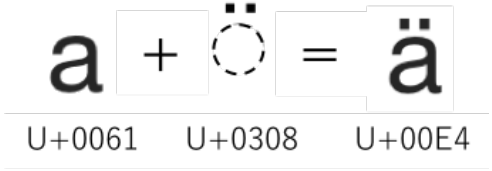
PRECIS Framework

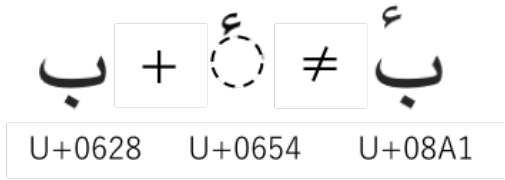
- **Stringprep**の課題を解決するために策定された国際化文字列を扱うための文字列処理の枠組み
- **Stringprep**からの主な変更点
  - 文字列変換処理の改善
    - **Unicode**正規化方式が選択可能
  - 特定のバージョンに依存しない文字列確認方法
    - **UCD**の情報に基づき利用可能・不可能な文字を分類
    - **Bidi**ルールの改善（**RFC 5893**に基づく確認が可能）
  - **2**つのサブクラスをサポート
    - **IdentifierClass**（識別子等）
    - **FreeformClass**（表示名，パスワード等）



IDNA2008/PRECIS Frameworkの課題

- **Unicode 7.0.0**以降への対応（現在最新の**Unicode**は**13.0.0**）
- **IANA**が**Uniocde 7.0**以降の**Uniocde**のバージョンの**IDNA2008**及び**PRECIS Framework**のプロパティ情報を更新しないとしたことにより，**IETF**で策定された主要な国際化技術が**Uniocde 7.0**以降の文字集合に対応できない問題が起きている
  - **Uniocde 7.0**にて新たに収録された**ARABIC LETTER BEH WITH HAMZA ABOVE (U+08A1)** が等価とみなすべき文字コードのプロパティに等価性を示す情報が記載されていない（**NFC**により合成されない）
  - **Unicode 13.0**でも問題が継続している





  - **Unicode 10.0**で戸籍統一文字や住民基本台帳ネットワーク文字等の行政システムで利用される文字はほぼ一通り収録された文字も利用できない

Unicode 7.0以降への対応提案(1/2)

1. **IETF**が策定する国際化技術の標準化推進として**IETF**の主要な国際化技術である**PRECIS Framework**の**Unicode 7.0**以降への対応提案

➢ **IANA PRECIS Derived Property Value registry**の更新が必要

- 現在登録されているのは**Unicode 6.3**のみ
- **IDNA2008**に関しては，**draft-faltstrom-unicode11**にて更新提案がされている

- **PRECIS Derived Property Value**を計算するプログラムの開発
- 上記プログラムを用い，**Unicode 7.0**から**Unicode 12.0**までの各**PRECIS Derived Property Value**を算出し，その差分を調査
- 調査結果をまとめた**draft-nemoto-precis-unicode12-00**を執筆し**IETF**に投稿
- **draft-nemoto-precis-unicode12-00**について関係者と共有

Unicode 7.0以降への対応提案(2/2)

draft-nemoto-precis-unicode12-00 (Informational Document)

- **PRECIS Framework**が**Unicode 12.0**に対応可能か，**Unicode 6.3**からの各**Unicode**(各**1,114,112**文字)の**PRECIS Derived Property Value**の差分について説明
  - **draft-klensin-idna-unicode-review**の提案に従い，**Unicode**のメジャーアップデートについてのみ比較
  - 改版時に追加された文字の多くは適切に**Derived Property Value**が割り当てられていることを説明
  - **Unicode 8.0**で追加された**SHARADA SANDHI MARK (U+111C9,**  **)**は，**Unicode 11.0**にてプロパティ情報に変更が生じ，**Unicode 8.0 - Unicode 10.0**までは**ID_DIS or FREE_PVAL**が割り当てられていたが，**Unicode 11.0**以降は**PVALID**となり，この特殊な変更について本**I-D**では，**IDNA2008**との整合性を考慮し，例外文字コードリストへの追加は行わないことを提案

Appendix E. Changes from Unicode 10.0.0 to Unicode 11.0.0

Changes from derived property value ID_DIS or FREE_PVAL to PVALID.

```
111C9        ; PVALID      # SHARADA SANDHI MARK
```

Changes from derived property value UNASSIGNED to either PVALID, DISALLOWED or ID_DIS or FREE_PVAL. Note: "ID_DIS or FREE_PVAL" is written as "FREE_PVAL" for convenience.

```
0560         ; PVALID      # ARMENIAN SMALL LETTER TURNED AYB
0588         ; PVALID      # ARMENIAN SMALL LETTER YI WITH STROKE
05EF         ; PVALID      # HEBREW YOD TRIANGLE
07FD         ; PVALID      # NKO DANTAYALAN
07FE..07FF   ; FREE_PVAL   # NKO DOROME SIGN..NKO TAMAN SIGN
08D3         ; PVALID      # ARABIC SMALL LOW WAW
09FE         ; PVALID      # BENGALI SANDHI MARK
0A76         ; FREE_PVAL   # GURMUKHI ABBREVIATION SIGN
0C04         ; PVALID      # TELUGU SIGN COMBINING ANUSVARA ABOVE
0C84         ; FREE_PVAL   # KANNADA SIGN SIDDHAM
1878         ; PVALID      # MONGOLIAN LETTER CHA WITH TWO DOTS
1C90..1CBA   ; PVALID      # GEORGIAN MTAVRULI CAPITAL LETTER AN..GEORGIA
1CBD..1CBF   ; PVALID      # GEORGIAN MTAVRULI CAPITAL LETTER AEN..GEORGI
2BBA..2BBC   ; FREE_PVAL   # OVERLAPPING WHITE SQUARES..OVERLAPPING BLACK
2BD3..2BEB   ; FREE_PVAL   # PLUTO FORM TWO..STAR WITH RIGHT HALF BLACK
2BF0..2BFE   ; FREE_PVAL   # ERIS FORM ONE..REVERSED RIGHT ANGLE
2E4A..2E4E   ; FREE_PVAL   # DOTTED SOLIDUS..PUNCTUS ELEVATUS MARK
312F         ; PVALID      # BOPOMOFO LETTER NN
9FEB..9FEF   ; PVALID      # <CJK Ideograph>..<CJK Ideograph>
A7AF         ; PVALID      # LATIN LETTER SMALL CAPITAL Q
A7B8..A7B9   ; PVALID      # LATIN CAPITAL LETTER U WITH STROKE..LATIN SM
A8FE..A8FF   ; PVALID      # DEVANAGARI LETTER AY..DEVANAGARI VOWEL SIGN
10A34..10A35; PVALID       # KHAROSHTHI LETTER TTTA..KHAROSHTHI LETTER VH
```

3.5. Changes between Unicode 10.0.0 and 11.0.0

Change in number of chacters in each category:

PVALID changed from 123,734 to 124,136 (+402)

UNASSIGNED changed from 837,775 to 837,091 (-684)

CONTEXTJ did not change, at 2

CONTEXTO did not change, at 25

DISALLOWED changed from 140,429 to 140,430 (+1)

ID_DIS or FREE_PVAL changed from 12,147 to 12,428 (+281)

TOTAL did not change, at 1,114,112

Code points that changed derived property value from other than UNASSIGNED: 1

As explained in Section 4.3 of draft-faltstrom-unicode12-00 [I-D.faltstrom-unicode12], there are some Unicode General Properties changed. Changes of properties for Georgian letters in the ranges U+10D0..U+10FA and U+10FD..U+10FF, ZANABAZAR SQUARE VOWEL SIGN AI (U+11A07) and SPHERICAL ANGLE OPENING UP (U+29A1) do not affect PRECIS calculation of the derived property values.

Change of SHARADA SANDHI MARK (U+111C9) added in Unicode 8.0.0 affects PRECIS calculation of the derived property values in

標準化推進に向けた課題

- 本課題について議論すべき適切な**WG**がない
  - 国際化技術に関して議論を行う**i18n-discuss ML**も閉じるべきか議論があり，閉じるべきでない旨を伝えた
- 国際化技術に関する文書のレビューを行える専門家が少なく，国際化技術に関する議論に遅れが生じている
  - **Internationalization Directorate**（**i18ndir**）
- **i18ndir**の専門家等に相談を行い，本**I-D**と**draft-faltstrom-unicode12**を統合することで標準化を進めることが可能か統合可能性を調査
  - **Unicode12.0**における**IDNA2008**と**PRECIS**の差分を調査し，**I-D**の改版を準備中
    - **IdentifierClass**で「**1,409**」文字， **FreeformClass**で「**14,103**」文字差分があった

国際化技術を利用するIoT技術の課題共有(1/2)

2. **IoT**サービスの情報資源参照時における利便性や安全性，相互運用性の向上に必要となる国際化技術の観点からみた課題の共有

- **dnssd WG**，**homenet WG**，**core WG**等で策定するプロトコルで，識別子に**UTF-8**が使用可能なものが存在する
  - **IDNA2008**や**PRECIS**への参照はない
  - 標準化された国際化技術を使用しないと独自実装等により意図しないサービスを参照してしまうことやサービスに到達できない可能性がある
- **W3C**でも**WoT**で**core WG**の**CoAP**が利用可能であることや**Thing Descriptions**を記述する際に国際化文字列を使用する際は**RFC8259**を参照する旨が記載
  - **WoT**では，国際化文字列を扱うための文字列処理等は定義されていないため**IETF**での課題が直接影響する

国際化技術を利用するIoT技術の課題共有(2/2)

- 国際化文字列の処理に課題があったUTF-8を識別子として利用するIoTサービスディスカバリ技術
  - **dnssd WG**（**RFC6763**，**draft-ietf-dnssd-hybrid**），**homenet WG**（**draft-ietf-homenet-simple-naming**），**core WG**（**draft-ietf-core-rd-dns-sd**）等で，**mDNS**（**RFC6762**）が使用する国際化技術の影響を受けている
  - 特に日本語の文字列処理における**Width mapping**の必要性や制御文字に関する問題がある
- 本課題について，国際化技術側の関係者等に情報提供を行なった
  - 参照先とする国際化技術である**IDNA2008**及び**PRESIC**の更新作業が完了していないことから，提案者らからの不用意な反発を避けるため，まずは国際化技術の専門家として**i18ndir**や**artart**の関係者に共有

まとめ

- 国際化技術の専門家に対して本調査を通じて得た課題，国際化技術の **Uniocde 7.0**以降の対応，に対して関心を得てもらうとともに，課題解決に向けた方針案について議論をおこなっている
  - それを解決するための公の議論の場が見つかっていないことが課題となっている
- **IoT**機器のサービスディスカバリに関しては，**DNS-SD**を利用した手法が検討されているが，**UTF-8**を識別子として利用する提案では，日本語を含む国際化文字列の処理手法の利便性及び安全性に関する課題が継続している
  - **IoT**機器のサービスディスカバリに関して，国際化技術の専門家以外にも課題を理解してもらう課題解決に向けた議論の場を醸成する必要がある
- **W3C**では，**IETF**が国際化を考慮してプロトコルを設計していることを前提としているため，**WoT**では**IETF**の課題が影響する可能性がある