White paper on Handling of Multiple Character Sets

Taking the Debate to the Logical Level through Market Practice Development

(Version 1.0, June 2010)

Note: Relevant regulations and any applicable legislation take precedence over white papers issued by this body. This paper represents an industry's effort to inform peers on relevant topics. The PMPG - or any of its Members - cannot be held responsible for any error in this white paper or any consequence thereof.

1 Introduction

The Payments Market Practice Group (PMPG) is an independent body of payments subject matter experts from Asia Pacific, Europe and North America. The mission of the PMPG is to:

- take stock of payments market practices across regions,
- discuss, explain, and document market practice issues, including possible commercial impact,
- recommend market practices, covering end-to-end transactions,
- propose best practice, business responsibilities and rules, message flows, consistent implementation of ISO messaging standards and exception definitions,
- ensure publication of recommended best practices,
- recommend payments market practices in response to changing compliance requirements

The PMPG provides a truly global forum to drive better market practices which, together with correct use of standards, will help in achieving full STP and improved customer service.

2 Introduction

The discussion of character sets in the area of cross border payments has never been easy. In short, attempts to introduce non-Latin characters in this space faced steep challenges. On the one hand, there is a requirement of clear cut standards and rules[1] to restrict character sets to basic Latin to cater for the limitations of bank legacy systems as to what a bank system can handle and what it can't. On the other hand, there is a growing, somewhat emotional frustration in the non-English speaking communities that requirements fulfilled in the domestic scene, are even hardly discussed when expanding into the cross border scene.

And yet, as internet based non-bank service providers proliferate the consumer markets throughout the world with no constraints on character sets, it seems to be the

[1] For example, please see SWIFT, "Frequently Asked Questions on character sets and languages in MT and MX free format fields" dated 20 February 2008.

appropriate time to revisit the issue of character sets. In this paper an attempt is made to set out the logical steps when considering the introduction of non-Latin characters in the area of cross border payments. It also suggests making the distinction between "standards" and "market practice", as well as the collaborative engagement of organizations in both domains, will be the driving force for further development of this issue. While our younger generation freely sends smiley faces and objects over their cell phones, for the sake of clarity, we would like to limit our discussion to characters (letters) used in the current business language.

This paper does not intend to go to the technical detail of the issue, but a simple glossary is attached at the end for readers with further interest.

3 Understanding the underlying technical constraints

Before we start debating the issue, let's take a step back and share a common understanding of why and how character sets become an issue.

The underlying constraint is simply that machines (business computers) can only process what they are programmed to process. As for characters, each character has a numerical value assigned to it that gets translated into a machine readable form consisting of zeros and ones. Where older generations of software are only capable of reading basic Latin and numeric characters, newer generations of computers are potentially capable of accepting any recognized character in the world.

The financial industry -especially the bank to bank space- belongs to the former where a significant portion of text is still stored and handled in legacy systems based on legacy encoding. On the other hand, newly developing regions such as China and Brazil use new information systems with modern dominant standards where the handling of non-Latin character sets is the norm. This standard is often referred to as *Unicode* and is natively implemented in recent technologies such as XML, Java programming and Linux operating systems. However, one should also take note that standards around character sets are not as stable and occasional revisions need to be factored into the discussion.

Since cross border payments involve multiple actors spanning across multiple geographies, implying the use of various generations of computers, the issue of character sets manifests itself as follows: if a character -not intended to be processed at a particular point within the payment chain- enters a computer system, the output may become unreadable or in the worst case, the payment application may crash. In other words, a character that the originating institution has passed on to the next institution in the payment chain, will eventually not be received in its intended form at the beneficiary or the beneficiary institution.

In the section that follows we present a five step approach to enable the transportation of information as intended by the originator throughout the end-to-end payment chain.

4 A Five Step Approach for inclusion of non-Latin characters in cross border payments

The following five points are proposed as the fundamental principles in dealing with the issue of character sets.

1. Tags or fields where non-Latin characters can be used must be agreed within the given community together with the additional characters that can be used
2. A character that is received must be able to be screened, processed and output in its intended form as required by the business
3. A character that has been received must be transported to the next party in the payment chain in a machine processable manner
4. Where a character that is received cannot be transported to the next party in the payment chain, the structures and guidelines articulating the responsibilities of translation or transliteration or transcription need to be developed and agreed
5. The above steps must be defined consistently for the same payment elements across different business domains

Step 1: Tags or fields where non-Latin characters can be used must be agreed within the given community, together with the additional characters that can be used.

English as a business language is a widely accepted "practice" and it is conceivable that not all tags / fields may require the use of non-Latin characters. Given the variety of actors in a payment chain, a certain degree of predictability is required, so an agreement would be required as to what extended character set the cross border community is willing to accept.

Since character sets are inseparable from "language", the agreement of tags / fields is important because for each tag / field different conventions may be defined around the content with regards to Step 3 of this proposal. For example, agree on structuring conventions for a "Name" or an "Address", in order to facilitate the understanding for correct transport of the data.

Step 2: A character that is received must be able to be screened, processed and output in its intended form as required by the business

Financial institutions acting as intermediary to a cross border payment might be required to screen the received payment and the relative data against a sanctions list. Where sanctions lists are currently in basic Latin it is conceivable that with the growing membership of the FATF, sanctions lists could be expressed in the original character sets, rendering the legal and compliance risk complex.

And while the business would process the characters as long as they are machine readable, the issue would always be the one of human readability and accurate interpretation for those parties requiring output in the form of reports, statements and advices.

Step 3: A character that has been received must be transported to the next party in the payment chain in a machine processable manner.

In the ideal world, a character accepted at the originating point of the end to end payment chain should be transported in its intended form to the end of the

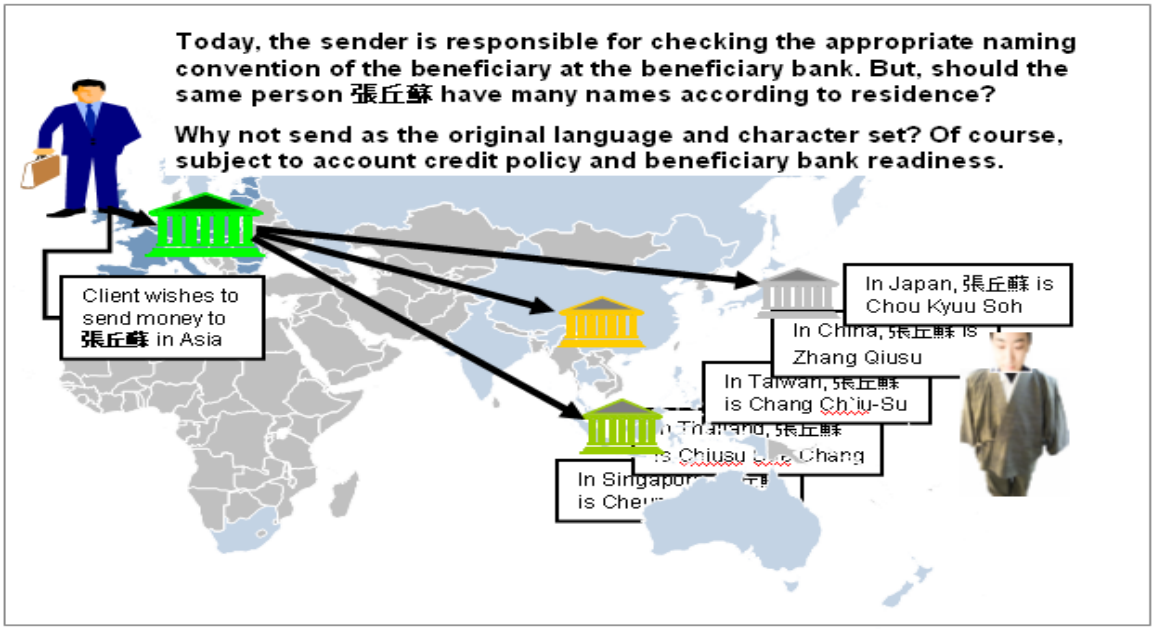
payment chain. However, given the current constraints, this may not be possible and certain "translation" techniques or "mapping dictionaries" require development in the form of bilateral agreements or as a community consensus. The point is that there should be an agreement within the community that characters that have been received must be transported to the next party and not be discarded.

Step 4. Where a character that is received cannot be transported to the next party in the payment chain, the structures and guidelines articulating the responsibilities of translation or transliteration or transcription need to be developed and agreed.

While there are many ways of expressing non-English language using basic Latin characters, there are limitations because language and characters come together.

Various techniques including translation, transliteration and transcription could be used to fit characters into basic Latin. When taking this approach, responsibilities of adopting such converting methods need to be clearly articulated and agreed upon. This would include the contractual responsibility with the relation to a customer as well as building convention and consensus of delineation of legal risk and responsibilities among the various intermediaries in a cross border payment. The SEPA extended character set practice addresses this through the combination of a SEPA conversion table and a set of principles which allows the originator to use local language characters in certain text fields and makes the receiver responsible for the conversion[2].

In addition, as shown in the diagram below, multiple transcription or transliteration for the same character is possible even within the same region. The illustration shows that the same Chinese character shown in the upper left corner is transcribed differently depending on country throughout Asia, which speaks for the introduction of original language characters in the cross border space.



Step 5. The above steps must be defined consistently for the same payment elements across different business domains

Since characters are a fundamental basis of a transaction, consistencies in the principles of dealing with character sets will benefit the participants of payment schemes, for example, the SEPA payment scheme or a US or Asian PACS variant. In a similar way, consistencies in the principles for handling a character set in a cross border payment or a payment related to a corporate action on a class of securities should also benefit the users of such schemes.

5 A collaborative approach to the character set issue

Since we cannot expect that all global banking systems become capable of accepting and processing all known characters immediately, a logical and realistic approach is required to accommodate the requirements mentioned before.

First, it is important to delineate standards issues from market practice issues. While market practice evolves into conventions and could be set as standard, the following is the current state of play as we see it.

1 Typical Standards Issues
   - The definition of a character set, the definition of transcription and transliteration schemes can all be seen as standards issues.
   - ISO20022 is based on Unicode using UTF-8 encoding which is a computing industry standard that allows for most known character sets.
   - The principles around the use of extensions within the ISO 20022 normal release cycle is a standards issue.

2 Typical Market Practice Issues:
   - Accepting English as the only business language and basic Latin as the only process-able character set is a typical market practice issue.
   - Defining the business requirement as well as articulating the business need of non-Latin characters is another typical market practice issue.
   - Discussions on responsibilities across various actors in the cross border payment chain as set out in Step 4 above are market practice issues.

Standards bodies and market practice groups should work together to structure and construct a collaborative framework towards the gradual adoption of non-Latin characters in cross border payments.

6 Proposed next steps and introduction of the PMPG

By working together with independent market practice groups such as the PMPG, there is a better prospect of capturing truly global and forward looking requirements that lead to better adoption of standards over time. It is expected that adoption and migration by complete communities will take time since the issue is by nature very technical and impacts the requirements on the underlying generations of systems. For this reason it is important to work out the principles and guidelines to set the direction for the industry to migrate to.

The proposed next steps would include the following, once the bodies for taking this initiative forward are agreed.

- Identify communities or chains requiring the use of non-Latin characters
- Define provisional end to end business requirements with reference to 5 steps
- Discuss and delineate market practice issues from standards issues
- Identify pilot communities / chains and support creation of business case
- Ensure feedback is provided for the benefit of further cases

7 Glossary of technical terms and introduction to standards

7.1 Definitions: Character Sets and Character Encoding

- Character Set: A very loose definition of a character set would be that it is a group of characters (often referred to as a repertoire of characters) that a system (or a community) chooses to support.
- Coded Character Set (or Code Page): A coded character set specifies how to represent a repertoire of characters using a number of non-negative integer codes called code points. It is a "character to number mapping table".
- Character Encoding Scheme: A character encoding scheme specifies how the integer code values should be mapped into an octet sequence (8 digits=8 bit= 1 byte binary code) suitable for machine processing. The number of octet sequences can be one or more.

7.2 Definitions: Language and Characters

- Cyrillization and Romanization of characters: A way of expressing the pronunciation of a certain language using Cyrillic characters or Roman alphabet. It is considered to be one form of transcription as shown below.
- Transcription of a language: The mapping of a pronunciation of one language to the best matching script of another language. Different characters with the same pronunciation tend to end up with the same transcribed character.
- Transliteration of a word: The precise mapping from one system of writing into another. Transliteration attempts to be exact, so if given appropriate tools, the original spelling can be reconstructed.

7.3 Standards

- Unicode: Unicode is a computing industry standard developed through the coordination by the Unicode Consortium, which allows the representation and manipulation of text expressed in most of the world's writing.
- Universal Character Set (UCS): UCS is an ISO/IEC 10646 standard and has been developed in conjunction with Unicode. In other words UCS and Unicode share their character repertoires.
- UTF (Unicode Transformation Format) -8: UTF-8 is the most commonly used encoding scheme for the Unicode text. (e.g. GB18030 is an encoding form for Unicode and defines the official character set of the Peoples Republic of China. GB18030 contains transformation rules to and from Unicode when needed. China declared in 2000 that any computer sold in its territory must support GB18030 and as a result banks in China use systems that support UTF-8.) (this actually means that the banks can use systems which supports Unicode if this system knows how to map GB18030/2005 and Unicode)