FOREIGN CHARACTER DOMAIN NAMES

WIPO -- GENEVA FEBRUARY 2001

"Foreign Character domain names arrived in 2000 under a shroud of confusion. The mists are now lifting, the market is taking off and they will be an important and relevant feature of TLD's now and into the future"

1. INTRODUCTION

PURPOSE

Foreign Character domain names (known as Multilingual domains to their friends) have been a long time coming, but now they've arrived. Despite the hubbub and hoopla surrounding the launch of Multilingual.COM domains in year 2000, myths and misunderstandings still shroud the appearance of this new addition to the Internet space.

The purpose of the paper is to explain in plain words what Multilingual domain names are, clarify some misunderstandings, shine a light into the arcane world of Internet standards, and signal the kinds of activities we are beginning to see now that Internet users are have begun to embrace Multilingual domain names.

CONTENT

It's a tough challenge to do justice to the Multilingual space in 15 minutes, but during this presentation, we should be able to cover the following aspects,

a. The domain name infrastructure may appear to be reliant on the "English language". However, its underlying "native tongue" is based upon ASCII (American Standard Code for Information Exchange), the ubiquitous encoding method for modern day computers. Multilingual solutions are not based on particular Languages or Countries either; rather they are based on language "Scripts" and "Character sets".

b. There are many local language solutions found on PC's today. Much of the technology is already deployed on your PC and used at set-up to "localise" or "countrify" the machine; i.e. to configure the language, the correct time zone, the dictionary used, the currency symbol, whether a "period" or "comma" is used as the numeric separator, et al. The ability to "Input" and "display" characters in a particular script is crucial.

c. The need to pass information in potentially different languages between 2 computers connected across the Internet gives rise to one major challenge: What is the "common denominator language" that allows both computers to communicate and understand each other, yet still cope with the huge diversity in languages world? The ability to "encode" and "translate" to and from that "common" form is crucial.

d. The challenge of finding that "common denominator" has been taken up by the IETF (Internet Engineering Task Force). Their standards process is focused specifically on establishing and improving mechanisms that ensure Multilingual domain names can be encoded, passed across the 'Net, and successfully translated at the receiving end. Whilst this may appear to be a big task, it has some focused goals and considerable progress has been made already.

e. Technical standards are only one (yet important) part of a much larger Multilingual picture. In parallel with the IETF task, domain Registries and Registrars are now gearing up for the introduction of multilingual domains into their name space. We will cover some Multilingual activities that you are likely to see going forwards.

In concluding, Multilingual domains are now an important addition to the domain name system and are burgeoning on the Internet landscape as we move forwards. Standards are always going to be an aspect requiring attention (for this or any other service over the 'Net), but they need to be placed and understood in the context of what they are focused on, and what results they are trying to achieve.

Foreign Character domain names are now being introduced by Registries and Registrars globally, and will need to feature on every TLD managers agenda going forwards.

2. SCRIPTS AND CHARACTER SETS

It is generally understood that the Domain Name System (DNS) is based upon the English Language. This is presumably because of the oft-quoted three letter TLD (Top Level Domain) strings. These strings are English acronyms, examples representing,

- com -- Commercial
- net – Network
- org – Organisation

Sign language for the Deaf

Signing was around as far back as Ancient Egypt, evolving as a language to allow deaf people to communicate.

In Ancient Greece, the killing for deaf people was a common practice, Aristotle commenting that those born deaf "*become senseless and incapable of reasoning*".

Perhaps "deafness" was a contributor to our ancient world's "digital divide" and that a universal language in the form of signing was a major step towards bridging that divide.

However, whilst traditional use may view these as English, Table A shows TLD strings from a different viewpoint, i.e. how do they look when interpreted in a range of different languages?

TABLE A: "ENGLISH" GTLD'S

LANGUAGE ENGLISH ACRONYM	.COM Commerce	.NET Network	.MIL Military	.PRO Professional
SPANISH	Dismal Common Ordinary		Thousand	
PORTUGUESE	With Upon		Thousand	Through
ESPARANTO			Thousand	Owing to
DUTCH		Beautiful		
AFRIKAAN		Alone Just		
SWEDISH			Mile	
DANISH			Friendly Kind	
LATIN				Instead of Before

Whilst Dot.com may stand for Commerce in the USA, how do Spaniards feel when registering names in Dot.Dismal"?

What about Denmark, they really have an army to be proud of, probably the most "friendly" "Dot.mil" known to man-kind ☺

And let us not ignore how the French might see things; in the recent ICANN meetings, the Board rejected "Dot.fin" TLD. This truly was "The End" ☹

Whilst the TLD strings are based upon a common **CHARACTER SET**, we've seen this does not translate into a common language. Characters are grouped together to form **SCRIPTS**, each defined Script drawing together all the Characters required for one, or more, languages. For instance, the Western European script includes all the characters required to represent the 12 languages shown in Table B. This includes,

- Numerals: 0 1 2 ...
- Individual letters: A B C ...
- Both upper & lower case: a b c ...
- All accents: grave, cedilla, acute, umlaut ...
- All punctuation symbols: < ! ? > ...

Each Script is typically given a reference, many being defined by the Institute of Standards and assigned an **ISO** number. They often are also known by names that are more common, a few are represented below in Table B. The characters that we see in the TLD Table A above are all taken from the **Western European Script** (also known as **LATIN 1**).

TABLE B: EXAMPLE SCRIPTS AND LANGUAGES COVERED

WESTERN EUROPEAN ISO 8859-1 (LATIN 1)	EASTERN EUROPEAN ISO 8859-2 (LATIN 2)	ARABIC ISO 8859-6
• Albanian • Basque • Catalan • Danish • Dutch • French • German • Italian • Portuguese • Rhaeto-Romanic • Spanish • Swedish	• Croatian • Czech • Hungarian • Polish • Romanian • Serbian • Slovak • Slovenian	• Arabic • Pakistani Urdu • Persian

The Domain name system was designed to operate with a very limited set of characters, these characters being part of the Latin 1 Script. This means that only 37 characters are currently allowed in the domain name, a very limited set of characters indeed given the number of different languages in the world and the variety of characters that exists within each one.

- Lower case: a to z
- Numbers: 0 to 9
- The hyphen: "-"

In order to enhance the system to allow Foreign Character domain names, the DNS needs to be expanded to,

- Remove the restriction to list names using only a single Script set, ASCII
- Permit the use of a range of different Script sets
- Define the Scripts that will be permitted
- Re-define any character set restrictions within each Script

Considerable activity is underway by the IETF standards group to see how to expand the number of Scripts & Characters to cope with the range of countries and languages used within each country. DNS expansion may, at first glance, appear related to countries and their languages.

However, this is not the prime focus in opening up the DNS to handle Multilingual domain names. **Scripts and Character sets**, not countries and their respective languages, provide the technical challenges that the standards groups need to focus on.

TABLE C: EXAMPLE COUNTRY AND SCRIPT MIX

COUNTRY	LANGUAGES SPOKEN	LATIN	CHINESE	INDIAN
SWITZERLAND	• English • French • German • Italian	**Y** **Y** **Y** **Y**		
SINGAPORE	• English • Chinese • Tamil	**Y**	**Y**	**Y**

3. FOREIGN CHARACTERS RUNNING ON STAND-ALONE PC'S

Having got this far and survived, you are half way home already; that is the end of the conceptual theory. From now on, we get a little more pragmatic.

Before we can consider the issues of running a Multilingual service over the Internet, we need to first isolate some specific issues with running Foreign languages on a stand-alone PC.

Addressing these issues is a pre-requisite prior to steeping up to the Internet "plate". So, let us image that we've just bought a brand new PC, are based in Japan, and want to operate it using the Chinese language; what do we need to do.

First up, we need to configure the machine to the Chinese environment. This is done on arrival, and requires the PC to be "countrified", i.e. configure settings for language country; date, time and currency formats, choice of dictionary etc.

Braille language for the Blind

In 1821, a soldier named Barbier visited a blind school in Paris, bringing his invention -- "night writing". This system was designed to allow soldiers to communicate along the networks of trenches at night without the need to talk and reveal their positions.

The coding was based upon 12 raised dots that when combined, translated to represent different sounds. Louis Braille simplified the system and thus the Braille coding system was invented.

As for "Night Writing", it proved too complex for the soldiers and was abandoned. Who said anything about "military intelligence".

Next, we need to think about how we can **enter information** into the PC in Chinese. To ease our task, we may have already purchased a Chinese keyboard, containing characters in marked up in the Han Script. Alternately we may only have an English keyboard, so we will need to load some software on the PC to map the English keyboard to Han Characters. Thus, in order to get anything in, we will need to have an **INPUT METHOD** available on the PC.

Being able to enter information is one thing, being able to **see it on the screen** is quite another. The process of displaying characters on the screen is called "rendering", so we need to make sure that we have access to software with an appropriate **METHOD OF RENDERING.** A "**gif Renderer**" is a Rendering program that uses a "gif" format to display pre-defined pictures of the characters on the screen.

Where the PC does not come with a suitable Input or rendering method, these programs are available as "plug-ins" to your PC. Commonly used ones are the NJSTAR plug-in for the Chinese language, and the NETEX plug-in for Hebrew.

The final thing that you need to do is look at the **Application software** to see whether it comes already using Chinese, or whether it is "configurable" to change its menus et al to operate using the Chinese language.

Summarising, in order to successfully enter information and display it on your PC in a local language, you need to ensure there is an appropriate **INPUT METHOD** and display **RENDERING** software running on that PC. These are pre-requisites that you need to get, before moving to the next stage, getting the PC to communicate with another.

4. COMMUNICATING OVER A NETWORK USING FOREIGN CHARACTERS

In this first step, we will assume that we have two identical computers sitting side-by-side. Our first task will be to "swap" the information using a Floppy disk (a CD Rom would be fine too here). A little later, we will see what impact there doing this via the Internet.

Before we begin, we need to ensure that both machines have appropriate software that includes an Input Method & Rendering capability. Imagine that we've input some Chinese to a word processing package, can see it displayed on the screen, and now want to pass this text to the other PC.

We first need to **save the information to disk**, and in doing so we will be **ENCODING** it. Encoding is the process whereby the information entered into the computer is converted into a series of zeros and ones (a binary string) for ease of storage.

Have stored the information, the floppy can now be transferred to the other PC where that **information can be read**, or **TRANSLATED**, by the other machine. Because both PC's are identical, we can assume that the Encoding & Translation methods are compatible, i.e. that we do not have to do anything special to **DETECT** the kind of encoding passed from the first machine to the second.

Esperanto, a Universal "intermediary" language?

In 1887, Dr L.L. Zamenhof introduced the Esperanto language He conceived it as a language that would allow people who spoke different native languages to communicate, whilst still retaining their own languages and cultural identity.

Today millions of people speak Esperanto, although it has not supplanted anyone's language, simply serving as a common second language.

In the context of Multilingualism, Esparanto was a great innovation, and was well ahead of its time. It reduced the different languages to a lowest common denominator, a feature of today's Multilingual solutions.

The Internet is a network of many types of computers each being configured in an indeterminate way. This means that when using Foreign characters on the Internet, all three process above are critical to achieve success, i.e. **ENCODING, DETECTION & TRANSLATION**. The IETF standards group has spent considerable time focusing on this particular aspect.

5. KEY ASPECTS OF THE IETF STANDARDS PROCESS

At this point, you may be wondering, ***"If software is already available to allow PC's to Input, Render, Encode and Translate using Foreign Characters, what is all the fuss about?"*** If so, you are already seeing through some of the "myths" of Multilingual. It is not (totally) an arcane black art; much of it is common, readily available, and not subject to the current IETF Multilingual standards process.

The current standards process primarily focuses on one particular aspect of the use of Multilingual over the Internet – the **USE OF THE MULTILINGUAL DOMAIN NAME**.

The **domain name is an "Internet parameter"**, a basic building block whose use features widely across many part of the Internet (e.g. Web, Email, FTP, etc).

Because of this, the standards process concentrates on just the domain name itself, not on the Applications using the name, nor the Domain Name System that manages those names.

The Standards group have had to grapple with a number of challenges, not the least being how to overlay a new "feature" of Multilingual domain names whilst leaving the existing Infrastructure intact (i.e. in engineer speak, "not broken").

Unicode, an encoding system for today's' computers

Computers store letters, numbers, and other characters by assigning a unique number to each one. Prior to the introduction of the Unicode system, there were literally hundreds of proprietary encoding systems around.

Many languages were not encoded, and some languages, such as English, required multiple coding systems to ensure that all characters were truly represented.

Unicode changes that, as it provides a unique number for every character, no matter what language and no matter what the computer platform.

In the Multilingual world, it is of tremendous utility and has been selected as the intermediary language to pre-package and funnel Multilingual strings into the "Name Prep" process.

In this regards, it is fair to say that it plays out a similar role to that envisioned by Dr. Zamenhof for Esperanto back in the 1800's

As part of core Internet infrastructure, the domain name has to-date been encoded using ASCII (back to section 1). To maintain compatibility with the existing infrastructure, they chose to select an Encoding system that converts Foreign Character domain names into an ASCII-like code: Conclusion,

- **OUTCOME 1: Choose an ASCII Compatible Environment (ACE) format as the "target" encoding** system over the Internet. [For the technically minded, improvements to ACE are being finessed right now with number of variants (RACE, LACE, BRACE); the nuances are subtle and best ignored by this paper]

Enforcing an ACE environment "Internet side" ensures compatibility with the current system. Standardising on one format also reduces the levels of complexity, thus leading to a simpler "Internet Side" solution. However, the "client side" of the network connects hundreds of millions of users operating scores of languages; the variety is extreme.

To solve this puzzle, the Standards group has selected Unicode (see break out box) as the intermediary "client side" encoding for Foreign Character domain names: Conclusion,

- **OUTCOME 2: Choose Unicode as the "Client side" intermediary encoding**. This simplifies the encoding routines by removing the infinite "variety" that exists.

Having chosen known environments for the Foreign Character Input and Domain name Output side of the equation, the group also focused its attention on the encoding/translation process to map between the two environments. This mapping process is known as Name Prep, and there is much discussion on how to refine the process to improve the range of Scripts and Characters recognized by this process.

- **OUTCOME 3: Define a NAME PREP process as the means of encoding and translating between both environments**. The Name Prep process contains rules that define how to handle special rules within languages as well as rules that permit/inhibit particular characters or strings.

These are the key decisions taken by the IETF standards group to permit the smooth introduction of Multilingual domain names onto the Internet. Whilst much work continues in refining these processes, the final solution for multilingual is now likely to change little going forwards.

One final comment needs to be made over the potential Impact of the above decisions. Given that Unicode is the mandatory intermediary input to the Multilingual process, it presupposes that each Client machine has their machine(s) operating in Unicode mode. This may require changes to browser switch settings and some education on the client side. The current process may also require a "plug-in" at the client side in the short term to help assist the process.

- **OUTCOME 4: The Client side solution set out above enforces some specific requirements on the Client's machine**. This approach could lead to a rise in customer enquiries because of changes to switch settings and possible use of a plug-in. The need for good Multilingual education, documentation, and support will become apparent.

Closing the penultimate section, the IETF Standards process has covered a lot of ground, and now has a clear idea of the required approach in the short term. The focus on domain name aspects only is much tighter than many might be imagined at the outset. Introduction of client side solutions may result in some confusion at the Client end, and will certainly raise the need for increased documentation and education into use of Multilingual domain names.

6. SO, ARE OTHERS MOVING AHEAD WITH MULTILINGUAL NOW, AND IF SO HOW?

We are already observing growth in the use of Multilingual domain names, particularly in those countries that use non-Latin character scripts, have high levels of non-English users on the Internet, and have higher levels of local language web pages already accessible over the net.

We are starting to see English web sites adding a simple local language contact page to provide some service to native speakers (parallels with introduction of the web back in the early 90's?). A number of these are forwarding their Foreign Character domain to their local language page, thus including a class of Internet users that would otherwise be disenfranchised.

Multilingual is really emerging as a complementary product to the traditionally Roman (ASCII) domain names.

From a Registry perspective, there has been increased activity over the last 6 months on assessing what the impact of Multilingual will be, and what might be a suitable response. It appears that the "Should we respond" question is being replaced by "HOW should we respond".

Registries are starting to look at the Drivers and Inhibitors of Multilingual in their name space, as well as the potential competitive impacts that are surfacing. Registries are also deliberating whether to mix Foreign Character domains at the third level, or whether they should create new Second Level domains (2LD's).

Policy questions are under discussion relating to how many languages, over what time frame, what "string" (if a 2LD) is appropriate, any changes to the charter, how do we moderate. More pragmatic issues are being churned over, such as can the billing systems cope, how do we print Multilingual invoices, how can we tell/correct "typos", and what does this mean to the languages spoken by our call center staff.

Finally we are seeing technology issues too, can our system cope, do we need to develop Multilingual "smarts", where do we get them from, how might this change our API to the Agent/Registrar/Reseller network we have in place, what impact is it for them.

The good news is that there is a considerable amount of planning and consideration going on, this augers well for further Multilingual growth.

ABOUT I-DNS.NET INTERNATIONAL

Headquartered in Palo Alto, CA, with offices in China, Japan, Korea and Singapore, i-DNS.net International commits to providing standards-based, standards-bearing multilingual Internet solutions that bridge the Digital Divide.

Backed by US-based General Atlantic Partners LLC, i-DNS.net contributes to technical and policy deliberations at Internet Engineering Task Force (IETF), Internet Corporation for Assigned Names & Numbers (ICANN), Asia Pacific Internet Association (APIA), Unicode Consortium, World Wide Web Consortium (W3C), Pacific Basin Economic Council (PBEC) and has affiliations with the Asia Pacific Internet Association (APIA), Asia Pacific Networking Group (APNG), Asia Pacific Top Level Domain (APTLD) and the Internet Society (ISOC).

Since its inception in 1999, i-DNS.net has worked with strategic partners like VeriSign Inc, Register.com, Melbourne IT, dotTV Corporation, eNIC Corporation, interQ Corporation, OnlineNIC and New Cyber International. Together with local in-country partners, the Company has launched its registration services across the globe in more than 30 languages.

Today, i-DNS.net continues to contribute its expertise as a technology developer, infrastructure builder and registry manager to industry and end-user alike.

Services

Multilingual Internet Technology and Solutions Provider

Develops the Internationalized Domain Name System (iDNS) to enable people to use their language of choice for their domain name.

Managed Registry Services

Operates the Registry for Full Multilingual Domain Names and offers one-stop Registry outsourcing solutions targeted at Registry Administrators

Professional Registrar Services

Develops solutions that augment the operations and business potential of the Registrar Community

General Enquiries: +65-332-0073
Email: info@i-DNS.net

Marketing Enquiries: +65-248-6188
Website: www.i-DNS.net

ABOUT THE AUTHOR

PATRICK O'BRIEN
V.P. & G.M. REGISTRY

PATRICK.O.BRIEN@I-DNS.NET

www.i-DNS.net

Patrick O'Brien is Vice President and General Manager for i-DNS.net International, a company specialising in Multilingual Enabling Technologies and Managed Registry Services. Prior to taking up this position late 1999, he was CEO at Domainz, the company responsible for managing the New Zealand domain name Registry. During his 4 years at the helm of Domainz, he was responsible for the start-up and profitable growth of this Internet business, providing the ***.nz*** space with one of the most respected management track records in the Registry community.

Before joining Domainz, Mr. O'Brien spent 7 years working in a range of senior management positions for New Zealand Telecom, ranging from Regional Director Asia Pacific, Manager of the Broadcasting business, through to G.M. Maritime Radio – with responsibility for New Zealand's distress & safety radio service to ships at sea.

His business skills have been supplemented with a solid UK educational foundation, gaining degree level qualifications in IT, and in Marketing, as well as an MBA majoring in finance. His early years were spent immersed in the UK IT industry, covering a broad spectrum from development through to operations, across large and small systems, and a number of sectors. The nine years spent with a professional firm of UK chartered accountants were pivotal in moving his early focus from one of technology issues, to one of business solutions.

Since 1998 Mr. O'Brien has had a close involvement with Internet Governance issues; providing submissions to the US Govt Green and White paper processes, a member of the Boston Working Group's successful submission to US Govt leading to improvement in ICANN's Foundation Articles, presenting a paper as part of the original WIPO consultation process in Sydney 1999, and an active member of the multifarious ccTLD constituencies.