Internationalized Domain Names in the Management of Cultural Heritage

Cary Karp
Museum Domain Management Association

Abstract: Realizing the full envisioned value of internationalized domain names is contingent upon the ubiquitous availability of IDN-aware software. The requisite development effort is likely to be driven by the anticipated response of the user community. The user and developer communities therefore need to reinforce each other's participation in this initiative if it is to have reasonable prospects of success.

This paper describes relevant factors that have become apparent during the introduction of IDN in the dot-museum (".museum") top-level domain. The holders of names in that domain have long been concerned with the provision of localized digital content. The ability similarly to localize the identifiers of the repositories in which this material is stored gives welcome means for calling attention to, rather than obscuring, the languages of the cultures within which it originated, or of the audiences for which it is intended. However shallow the semantic value of domain names may be, they can indicate linguistic identity in a manner that is of clear interest to agencies such as museums, which are increasingly concerned with the role of language in cultural identity, and the preservation of language diversity.

Unicode is fundamental to all aspects of this process and this paper articulates needs and priorities in cultural action that may, in turn, be of use as the priorities of software developers are set. The paper is intended to further productive dialogue between these communities both to maximize the return from software development, and to ensure the greatest possible societal benefit from the availability of IDN as a banner of culture in the digital realm.

Introduction

The number of languages in which cultural institutions present material in on-line venues is constantly increasing. This development is driven from at least two different perspectives. The first is the translation of material in the primary language(s) used, for example on an institution's Web site, into one or more additional languages. The second is the harnessing of the Internet's potential for the distribution of material in languages that do not normally appear in other widespread publication media. In the former case, the intention is likely to be expanding the audience that can access an expression of metropolitan cultural activity, with the secondary languages selected on the basis of the size and mobility of their readerships. In the latter case, the activity may be conducted by, or on behalf of, peoples calling attention to small or diminishing cultures. In this situation, language is a primary attribute of those cultures and is their banner when addressing an audience that is targeted because of the role it can play in ensuring the survival of those cultures.

The preservation of intangible cultural heritage is the focus of considerable current attention. Language is a basic element of that material, whether as its primary substance, as part of performance events, as a vehicle for the transmission of traditional knowledge, or by being the key to understanding action that does not have an explicit linguistic component. There is a vital bridge here between the natural historical and cultural historical facets of the activity of institutions that deal with both, but traditionally see them as separate concerns. The manifold ways of conceptualizing and describing the world that are reflected in language are as relevant to the study of the way different cultures perceive the experience of life, as orally transmitted traditional knowledge of local fauna may be to the protection of endangered species, or similar knowledge of the medicinal properties of indigenous plants can be to the preservation of our own species.

It should be clearly within the mandate of an institution concerned with the management of cultural heritage to recognize its role in ecological contexts as it focuses on intangible actions and property. It should be similarly clear that the way in which such institutions broaden the linguistic horizons of their Internet presences can contribute to the preservation of the language diversity that fundamentally underlies more expressions of human knowledge and culture than we may yet have fully recognized.

The discussion of viable means for proceeding also needs to consider ways in which metadata and external resource identifiers can be used to highlight, rather than obscure, the language in which the material being presented originates. The technologies needed for the localization of content are reasonably well understood and development initiatives are underway in many contexts. One such action is the application of internationalized domain names (IDN) in enabling document repositories to be identified using the same repertoire of languages and scripts as is available for the material that they contain.

The value of measures such as this is often assessed in terms of anticipated enhancement of the user experience. The response thus far of name holders in the dot-museum (".museum") top-level domain to the availability of IDN suggests that the mere appearance of a language from the broader spectrum in a domain identifier, has a galvanizing effect that outweighs concern with many of the other factors that limit the current utility of IDN. This aspect of localization also appears likely to provide incentive to the contribution of material to the user community that otherwise would not have been forthcoming.

The Domain Name System

The Domain Name System (DNS) is the authoritative source of data associating the names of Internet hosts with their numerical addresses. It was devised to replace a centrally maintained reference table with a distributed database as the former mechanism approached the limit of its scalability [1]. The new system introduced a hierarchically structured namespace, branching from a root node into a number of top-level domains (TLDs). These divide into a larger number of second-level domains, which split further into as many subdomains and lower levels as may be required, constrained only by the maximum permissible length of a fully qualified domain name [2]. The nodal structure of the DNS follows administrative boundaries independently of network topology. A multi-sited enterprise can operate within a single domain regardless of the physical location of the facilities that it connects to the Internet.

Domain names were initially intended to be mnemonic alternatives to numeric addresses. They were 'protocol elements' and not supposed either to appear or be used as words or phrases. A major conceptual shift was precipitated by the advent of the graphic Web browser and its role in the growing use of the Internet as a platform for commercial activity. The address line in a browser exposed domain names and URLs to users in a manner that had not been intended by the developers of either protocol. At the same time, modems of sufficiently high speed were becoming available for Internet access to be gaining rapid inroads into the general consumer environment. Business sites were accessed by invocation of their domain names, which were therefore soon perceived in terms of brand value. The acquisition of names desirable in this respect quickly became aggressively competitive for both commercial and non-commercial enterprises. Trademarks began appearing in domain names, which were also treated as intellectual property in other regards. Speculators hoarded attractive names and the need for defensive registration became a major concern. Dictionary words abounded throughout,

introducing a semantic element into the DNS that extended its literal component into a role well beyond that of simple mnemonic convenience.

This process may inadvertently have been seeded by the use of clipped forms of English words to designate the generic TLDs <com, edu, gov, int, mil, net, org>. This also applies to the similarly derived two-letter abbreviations used in the parallel but far longer list of so-called country-code TLDs [3]. There is obvious mnemonic value in using the first three letters of a word to denote activity designated by that word, or in designating a country by a two-letter abbreviation of its English name (as many, but not all, such codes were derived). The explicit language attribute of labels created in this manner is, however, equally obvious. Whatever effect this may actually have had, 'dot-com' did become a full-fledged dictionary word in its own right, well known to people far outside the networking community, and of global economic significance. Together with the growing intellectual property interests, this gave rise to a fundamental reconsideration of the future direction in which the DNS should be developed. In 1996, the co-author of the original list of TLDs, Jon Postel, proposed accommodating changing needs and conditions through the establishment of a large number of additional generic TLDs [4].

Before discussing the ensuing action, one further language-related matter will be considered. The technical specifications of the DNS permit any of the printable characters in the Unicode Basic Latin code chart to be used in domain names, but related earlier specifications that restrict the characters that may appear in host names are still commonly applied [5]. These are limited to the twenty-six letters in that chart <A-Z, a-z>, the ten digits <0-9>, and the hyphen-minus < - >, referred to as the LDH characters. The general practice of TLD operators is to restrict all domain name labels to them. There are additional rules about the positions in which digits and the hyphen-minus may appear but nothing is said about the semantic aspects of the way in which domain labels can be constructed [6]. If a seemingly cryptic sequence of characters is appropriate for a given application, the DNS readily permits its use. Where explicit words and phrases are desired, they can appear with equal ease.

The operator of a domain may add further policy constraint to this, including both prescriptive and proscriptive naming conventions. These cannot, however, transcend the limitations imposed by the permitted character repertoire. If the LDH array provides a full orthographic basis for names in all languages toward which a domain's policies are geared, little more needs be said. If, however, the representation of target languages requires a larger repertoire, additional means will be needed for providing it. This is not solely a matter of enabling the appearance of explicit lexical items in domain names. If there are situations where Latin letters might reasonably appear in labels that are neither intended nor likely to be construed as words — abbreviations, initialisms, codes, etc. — there are equivalent situations where characters from other scripts are needed.

This is a straightforward internationalization issue and means for dealing with it have been considered in a variety of contexts. Relevant initiatives have been undertaken from differing perspectives but the results have yet to be fully integrated. One fundamental contribution is the specification for *Internationalizing Domain Names in Applications (IDNA)* [7], finalized by the Internet Engineering Task Force (IETF) in 2003, with the Internet Corporation for Assigned Names and Numbers (ICANN) posting corresponding *Guidelines for the Implementation of Internationalized Domain Names* in June of the same year [8]. Several TLDs had previously introduced comparable service on a proprietary basis and were

immediately able to proceed under the newly formalized terms (although the conversion of names registered in the previous manner was not an altogether trivial matter).

Issues in IDN deployment

A growing number of both generic and country TLDs are now accepting IDN registration in accordance with the ICANN Guidelines. One of the central requirements is that a registry base its IDN polices on language rather than script. The intention is to reduce the potential for confusion that would arise from the simultaneous availability of characters with different code points but glyphs with closely similar appearance. This potential already existed and was recognized in the LDH repertoire, for example, in the similarities between <1 \u0031> and <l \u006C>, or <0 \u0030> and <O \u004F>. It would, however, become significantly greater if scripts with overlapping glyphs such as Latin, Cyrillic, and Greek could be gratuitously mixed, with similar increase in potential for malicious exploitation:

Latin	Cyrillic	Greek	
<A \u0041>	<A \u0410>	<A \u0391>	
<E \u0045>	<E \u0415>	<E \u0395>	
<O \u004F>	<O \u041E>	<O \u039F>	
<a \u0061>	<a \u0430>		
<e \u0065>	<e \u0435>		
<p \u0070>	<p \u0440>		etc.

On first consideration, language-based IDN policies should be relatively easy to implement in a country TLD. There is normally a limited number of languages with formal status in any given country, toward which the national TLD registry could direct its IDN support. Language-based policies can also accommodate situations where multiple scripts are used for a single language, either as alternatives or intermingled. Although a categorical stricture on the appearance of multiple scripts cannot realistically be included in such policies, specific terms for multi-script labels can often be stated with relative ease. The situation is more problematic where a national registry deliberately wishes to support a range of languages in a broader political or cultural context and explicitly wants to identify that aggregate as such. A country belonging to a multinational union might, for example, support all of the languages used in that union, regardless of their official status in any single country, and deem it inappropriate to indicate any priority or bias by listing the component languages by name. In a situation such as this, policies based on the available character repertoire will be necessary, and the issues inherent in mixed scripts might legitimately be regarded as a secondary consideration.

The globally oriented generic TLDs lack any natural basis for association with specific languages. The range of languages that may be needed by the holders of names in these domains cannot easily be foreseen. Although there are compelling reasons for IDN support to be introduced in a generic TLD on a language-by-language basis, at some point the number of languages may become large enough to require policies treating them in aggregate. Reference to the Unicode standard will likely prove an inevitable component of such policies, as will the many separate technical reports that are appearing with greater or lesser normative intent. It is also reasonable to expect future versions of the ICANN Guidelines to recognize situations where the Unicode code charts can provide an appropriate basis for general IDN policies, either directly or by reference to them in the definition of scripts or “sets of languages”.

The essential feature of IDNA is that it makes no change to the DNS, itself, with registered domain names still being restricted to the LDH repertoire. Instead, IDNA specifies means for using LDH characters to represent a far larger array taken from the Unicode repertoire, through so-called ASCII Compatible Encoding (ACE). The encoded form of a name is not intended to be exposed directly to users, who normally see and use names containing IDN characters in their expected Unicode form. The corresponding encoded representation is termed Punycode. This is indicated by the characters 'xn--' in the first four positions in a label, the remainder of which is the output of the Punycode algorithm. In illustration, the encoded form of <lättöl> is <xn--lttl-loa4i>, and <ψάρι> is <xn--hxa0ax0a> [9].

All encoding and decoding is performed transparently by applications software. Given the relatively recent advent of this procedure, it is to be expected that there is as yet only a limited range of applications with corresponding functionality. Until IDN-aware working environments can be taken for granted, there will be a fledgling aspect to the endeavor. One consequence of this is that there are some situations where users may unintentionally see Punycode, and others where domain names need to be communicated openly in that format. Although the first of these situations may be a benign growing pain, the other increases the risk of confusion with potentially substantive consequence. This was one of the initial reasons for expecting Punycode to be veiled, although factors suggesting the untenability of doing so are becoming increasingly apparent.

IDN has probably already generated enough momentum in the domain name market to ensure its on-going presence. Its potential for serving purposes beyond that basic availability deserves closer scrutiny. The material that follows will illustrate one applications arena in which IDN can serve a clear and valuable societal role. As a preliminary to its detailed consideration, basic concepts in the management of cultural heritage will be discussed.

Cultural heritage

Culture is a complex aggregate of the attributes of human existence and resists terse general definition. It includes modes of behavior, belief, communication, creative expression, customs, kinship, and knowledge. There are awe-inspiring numbers of different systems but the variation among them is not continuous. Modally concentrated sets of cultural attributes can be associated with distinct social groups. The viability of a group is largely contingent upon its ability to transmit acquired knowledge to succeeding generations.

Heritage is a term commonly used to designate the body of information and material conveyed in this manner. The word is sometimes regarded as holding undesirable connotations of action completed in the past. This may be avoided by qualifying it as living heritage. Further distinction is made between cultural and natural heritage. In contexts where human interaction with the natural environment is being considered, separate reference may also be made to scientific heritage. The additional qualifiers, tangible and intangible, appear in many of these frames of reference. The term cultural property is used as a synonym for cultural heritage to permit convenient subdivision into fixed property (buildings, monuments and sites), moveable property (such as utilitarian implements and decorative objects), and intangible property (for example, orally transmitted knowledge and performance traditions).

The acquisition, maintenance and transmission of the body of knowledge, traditions and property regarded as heritage in the preceding senses, are profoundly dependent on language,

which is therefore often a primary identifier of cultural modality. The interdependence of linguistic diversity and cultural diversity has long been recognized. Attention has more recently been called to an equivalent relationship between language diversity and biological diversity. The accepted contraction 'biodiversity' has been re-expanded to 'biocultural diversity' as a designation for this new focus of study. It embraces a multiplicity of interactions internal to a community, between communities, and between people and the natural environment, thus spanning a range of academic disciplines on both sides of the traditional divide between scientific and cultural research, with language being a pervasive concern of them all.

Museums and the .museum TLD

The various facets of cultural, ecological, scientific, and language heritage and diversity, regardless of how they are combined or labeled, map strongly into the domain of museum activity. The International Council of Museums (ICOM) [10] provides the following definition of its institutional constituency:

> *A museum is a non-profit making permanent institution in the service of society and of its development, open to the public, which acquires, conserves, researches, communicates and exhibits, for purposes of study, education and enjoyment, the tangible and intangible evidence of people and their environment.*

This definition is subject to on-going modification as the museum profession assesses the dynamics of its societal role, carefully heeding the perceptions of the audiences that museums address. This also involves dialogue with groups that ascribe little relevance to institutional museums as a means for propagating knowledge about themselves and their cultures. Attempts at rectifying the latter situation have most frequently been directed toward geographic regions with low museum density, highlighting the potential of museums in bringing attention to dwindling and vital cultures, alike. Museums do this in several ways. The collection, interpretation, and dissemination of information remain prominent activities, but the traditional focus on material artifacts rather than on the dynamics of the situations in which they were produced, is often precisely what causes museums to be seen as irrelevant. Recent emphasis has therefore been placed on the role museums can play in enabling the voices of cultures to be conveyed directly from their sources to the intended audiences, with as little imposition of external values as possible. In the best of cases, this will solely be a matter of providing requisite infrastructure and facilitating its use. Where the bearers of a culture are fully occupied with ensuring their subsistence, museums will need to mediate more actively in the spreading of information about the imperiled situation, falling at least to some extent back on their traditional role. Finally, where attrition is unlikely to be reversed, the museum can strive to ensure a lasting record of that culture.

Areas of particular concern are also charted by UNESCO, which has been calling ever-clearer attention to the need for cultural institutions to focus on intangible heritage. This explicitly includes the utilization of the Internet in the maintenance of language diversity and for empowering smaller voices so that they can be heard without geographic confinement. At its General Conference in 2003, UNESCO passed a *Recommendation concerning the Promotion and Use of Multilingualism and Universal Access to Cyberspace* [11], which prescribes a detailed series of actions toward the *Development of multilingual content and systems* and

Facilitating access to networks and services. The fourteenth of these twenty-five recommendations is:

> *Member States and international organizations should promote appropriate partnerships in the management of domain names, including multilingual domain names.*

There is significant difference in opinion among museum professionals about the extent to which cyberspace can be seen as a cultural platform in its own right, and museums are extending their concern beyond the realm of physical artifacts in heterogeneous manners. ICOM was, nonetheless, an early participant in the discussion of new TLDs that was initiated in 1996. The story of how that interest resulted in the creation of the .museum TLD has been told elsewhere [12]. The fundamental premise in arguing the need for this domain was that Internet users would derive clear benefit from a segment of the TLD namespace being reserved for the use of museums as defined by the museum community, itself. This would enable a user uncertain about how to recognize the difference between a bona fide museum and a deceptive pretender, to verify the origin of material on the basis of the domain in which it originated. Conversely, this would provide museums with means for conveying trust, intermediate between attributes conveyed in-line with content, and formal third-party certification. The agency responsible for enforcing the threshold requirements for inclusion in .museum can, in fact, also be seen as a trust provider.

In the framework devised by ICANN during the establishment of seven new generic TLDs beginning in 2001 (the upshot of the 1996 proposal), the policy and eligibility bases for a domain that is restricted in this manner are specified in a charter. The .museum charter [13] stipulates that registration in the domain will be "granted only to entities that are museums [as defined by ICOM], professional associations of museums, or individuals who are professional museum workers [also as defined by ICOM]" with the policy maintenance agency being able to "extend this definition to cover other entities that acquire, conserve, and communicate or exhibit evidence of people or their environment, if petitioned to do so by a recognized professional organization within the museum community." (This proviso was added in recognition of the need to accommodate conditions changing at a rate that easily might require quicker response than is allowed by the three-year minimum period needed for the formal modification of the core ICOM definition.)

IDN in .museum

With the addition of .museum to the authoritative root of the DNS in October 2001, a well-defined platform for cultural activity was created on the Internet. In the doing, agencies within the museum community established a "partnership in the management of domain names" well before UNESCO issued its recommendation for undertaking such action. There was clear thought from the outset, of expanding this partnership through the creation of additional TLDs for adjacent segments of the heritage management sector such as .archive and .library (whatever the actual labels might ultimately be), with their policy foundations provided by the corresponding sectoral NGOs or in other globally representative collaborative action.

The label <museum> was also the first non-colloquial lexical item to be introduced into the TLD vocabulary [14]. Its semantic intent is both intuitive and clearly stated in the .museum charter, and direct lexical equivalents exist in all of the languages that are likely to figure in the

deployment of IDN. This provides an objective basis for the inclusion of .museum in any implementation program that might be developed to support what is often termed 'full IDN' — names that include Unicode characters external to the LDH repertoire on all levels including the TLD label. Preliminary action toward that end has already been undertaken in several country domains where the alternate language representations of the TLD label are obvious, and it is fully in keeping with the UNESCO recommendation for .museum to be cultivated similarly. Other generic TLDs also have plans for their multilanguage representation but the semantic clarity of the .museum name and the commonality of interest of its constituency suggest that it could serve well as a pilot case.

Whatever the course toward full IDN may be (with daunting policy, political, economic, and technical issues still to be overcome), accommodating the number of languages already manifested in burgeoning museum community response to the availability of IDN requires either that some languages be deferred to the advantage of others, or that extensions to the ICANN guidelines be devised. Selectively braking the gathering momentum would be regrettable in any case; all the more so if there is any truth to the contention that IDN can be a useful device in the urgent matter of reversing cultural attrition. Fortunately, there are factors inherent in the operation of .museum that allow its cultural objectives to be pursued with uncompromised attention to the overriding need for ensuing the stable operation of the DNS.

The most significant concerns about the introduction of IDN relate to its potential for causing confusion of the type already described. In high-volume registration situations where there can be significant real-time contention for names that are available without restriction, the registration process requires extensive algorithmic support. Adapting this to IDN without injecting disproportionate possibility for abuse is a significant technical challenge. There is, indeed, every reason to proceed in cautious increments, carefully monitoring the effects of the introduction of each successive language [15].

Regardless of the extent to which .museum registration procedures can be automated, there are no realistically foreseeable means that will eliminate need for sentient review of such things as an applicant's eligibility, or the conformity of a requested name to the domain's naming conventions [16]. The requirement, for example, that "every name registered in .museum must be clearly and recognizably derived from the name by which the entity to which it is assigned is otherwise widely known", immediately limits the constructs that may appear as labels. The ability to extend this to languages that cannot be represented with the LDH repertoire obviously requires commensurately greater linguistic expertise at the point of review.

There are further issues in the review process that require more detailed understanding of local conditions than can possibly be concentrated in a single administrative office. The established means for dealing with this are by referring uncertain requests to the ICOM national committee with the greatest familiarity with relevant local circumstances. There are currently 116 such committees, working in a wide range of languages and easily capable of ensuring that the manner in which these languages appear in .museum is fully compliant with accepted orthographic practice and other relevant linguistic detail. The character repertoire can therefore be expanded in direct response to the expressed interest of the museum community without jeopardizing the ability of the user community to trust the provenance of a resource bearing a .museum IDN. Any risk that might, nonetheless, be feared in consequence of the support of a large number of code points is offset by other restrictions of the .museum namespace, most particularly on the vocabulary that may appear in it.

The ICOM national committees may be used to illustrate two further points about IDN. The reader is reminded that the present discussion is lopsidedly focused on museum activity solely because .museum is the only TLD dedicated to the heritage management community, with the ICOM application being only one of several that could be described here. Anything that is demonstrated below will be immediately transportable into any other TLD that might appear in what is termed the 'heritage cluster'.

Impact

Organizations with broadly multilingual international constituencies often have difficulty in determining which languages to use for their official business. If nothing else, the expense of translation and multilanguage publication will invariably be a key concern. For many decades, ICOM conducted its central activity in the official languages English and French and recently added Spanish to that list. The national committees have, however, always conducted business in their own languages. As ICOM established its presence on the Internet, the language issue acquired a new dimension. The ease with which the national committees could provide localized content on their own Web sites strained the effort of developing a coherent organization-wide identity. The availability of the domain icom.museum provided a welcome degree of additional focus over the previous icom.org, and the next challenge was to convey a similar indication of the cultural substrata of the parent organization. Every committee had a name in the initial two official languages but the orthographic requirements of the French version left only the English one for use in the consistent naming of subdomains. This gave the intuitive general form 'committeename.icom.museum'. It also provided each committee with a persistent domain identifier, eliminating need for public concern with tracking changes to Web site and e-mail addresses as host facilities inevitably moved at regular intervals together with administrative responsibility for the committee.

The utility of this procedure was widely appreciated but there were varying degrees of enthusiasm about the restriction to English identifiers, especially from committees that did not maintain English language content on the designated platform. The availability of IDN promised swift rectification of this difficulty, and localized LDH committee names were released at the same time:

http://österreich.icom.museum
http://českárepublika.icom.museum
http://deutschland.icom.museum
http://ελλάδα.icom.museum
http://magyarország.icom.museum
http://ישראל.icom.museum
http://日本.icom.museum
http://한국.icom.museum
http://méxico.icom.museum
http://россия.ком.museum
http://slovenija.icom.museum
http://españa.icom.museum

A more extensive list including the scripts needed for native representations of all the national committee names is located at http://icom.museum/idn/natcoms.html [17]. When examining these lists, it should be noted that almost all of the ICOM committees use the Latin representation of the organization's acronym, and its juxtaposition with the Latin .museum is taken as equally devoid of language bias. Corporate identity is provided on the first two levels, and local identity is conveyed on the third. This branches further into the cultural realms identified by the languages used to localize the domain names, with the committees providing entry into a distributed body of cultural material that extends far beyond ICOM's administrative borders.

The .museum label provides one important signpost on this path. An IDN label can provide at least two others. The one follows directly from the indication of language, with all its extended semantic potential. The other is through the cultural association that both the language and script may have, independently of any direct meaning the label is intended to convey. Paradoxically, languages that use the LDH repertoire may be at a disadvantage to those where the IDN label can more readily be harnessed toward the preceding two ends. The indication of 'something special' conveyed by the mere appearance of an IDN, lacks counterpart in the workaday pre-IDN environment. It will take quite some time before the novelty value wears off, and more stable metaphors are certain to supplant it in the interim.

In as focused a context as that of .museum, ways in which this can be channeled are obvious. This may be illustrated by ICOM committee response to the release of localized names to them. Many that previously had felt themselves to be adequately served with no more elaborate a Web presence than the single English language page provided to them centrally (examples of which will easily be found in the extended list referenced above) undertook immediate action to establish more elaborate autonomous Web sites. Where these are multilingual, the IDN label is used to identify the localized document tree, and the LDH label for the English counterpart. For example, the Japanese URL in the preceding list leads to a Japanese language site, and http://japan.icom.museum/ to a corresponding site in English.

An unexpected additional effect was an inversion of the primacy of the languages in the ICOM hierarchy. This is a corollary to the paradox mentioned above, with graphemically distinctive languages being most easily able to garner visibility through the use of IDN, and English only minimally endowed with such potential, with French and Spanish dangling by their diacritics somewhere in between. The smaller languages used within the organization, which never would have come into consideration for official central use, acquired privileged status in the new context. This, in turn, gave impetus to additional Internet-based activity that would otherwise have been far longer in the coming.

The extent to which a script is shared indicates nothing about the sizes of the individual languages that use it, but immediate potential was seen for the application of IDN in bringing peoples and cultures associated with minority and indigenous languages into a significantly more visible position on the global stage than they would be likely to attain in any other manner. Further potential was seen in this for increasing the realization that museums are relevant in contexts where that had not yet been fully recognized. A myriad of issues needs to be addressed before the fuller practical extent of this potential can be determined. At this juncture it may suffice to say that the fact, alone, that IDN is extending the scope of museum concern with the role of language in the maintenance of cultural heritage is a worthwhile result, regardless of the form that the underlying technical facility ultimately takes.

Museum action toward increasing the appreciation of their relevance includes both dealing with agencies tending cultural property, and efforts designed to increase the public understanding of museums and the role they can play in society. At least one key aspect of the utility of IDN in these contexts is not specific to museums. Presenting the Internet to people who do not use it, obviously requires the dissemination of information through other communications media. The popular press frequently devotes space to precisely such effort. Any such publication in a language that does not use the Latin alphabet is less likely to achieve the intended goal if such things as potentially interesting Web sites are presented by their LDH names, than would be the case if IDN were available.

This extrapolates into contexts where the basic issue is making the Internet available to large new groups. Here the off-line message may be directed to governmental agencies, and the clear indication of cultural relevance that can be conveyed via IDN can be a compelling argument in itself, regardless of the language in which the political discussion is conducted. Similar emphasis may be placed on the cohesive effect the cultural attributes signaled by IDN may have for a diaspora in the maintenance of its cultural bonds.

Work still to be done

Further scenarios can be sketched illustrating the positive implications of IDN in the cultural arena. Doing so may, however, risk leading too far into a recursive presentation of the reasons why IDN was developed in the first place. Instead, the rosy picture shown thus far will be tempered by considering some challenges that remain before many of the goals delineated above can be fully realized. A detailed discussion of technical hurdles lies outside the scope of this paper, but a brief outline of the way they appear from the cultural horizon may be worth including, nonetheless. Each of the following bullet points comments on a single issue without any indication of priority.

- IDNA is currently locked to Unicode version 3.2. Unless it is updated, scripts that become available in later Unicode versions cannot be used for IDN. There is clear likelihood of the diffusion of IDN into the cultural contexts described here both resulting in such need and giving rise to expressions of interest in the inclusion of additional scripts. Indeed, IDN provides incentive for such action and thwarting it would be counter to interests that are not restricted to the cultural sector. The maintainers of the underlying normative instruments should therefore be particularly mindful of the need for acting in tandem to facilitate this process.

- Case folding can result in nonstandard orthography for words containing letters that do not have both uppercase and lowercase forms. Two that have already been noted as problematic are the GREEK SMALL LETTER FINAL SIGMA and LATIN SMALL LETTER SHARP S, neither of which can appear in Stringprep output. Further difficulty is caused by established alternate orthographies permitting the replacement of a permissible character by one that is either rejected by Stringprep, or fails at some other point. An example of this is the HEBREW PUNCTUATION GERSHAYIM <″>, which may prove necessary in IDN contexts (indicating, for example, that a label is an acronym) but does not appear on a standard Hebrew keyboard, instead being represented by a quotation mark <”>. People who understand the underlying issues need to become involved in the process of identifying and correcting problems of this type in the shortest possible time.

- IDN is exclusively intended for use in domain names. Although document repositories can easily be identified with nothing more, the individual items that they contain require the additional indication of file names. The recent finalization of a standard for *Internationalized Resource Identifiers (IRI)* [18] enables the localization of all components but also extends the scope of the development endeavor. The effect this will have on the rate at which IDN diffuses through the heritage management community cannot currently be assessed.

- The domain name in an e-mail address can include IDN characters but this does not extend to the mailbox name to the left of the <@>. IDN-aware e-mail agents can thus not fully meet what are likely to be common user expectations. Further difficulty may be caused by a non-compliant user agent displaying a sender's address in its Punycode form (with risk for serious damage if this were done in the local part), or when auto-replying to an address with IDN characters displayed in Unicode. Unexpected interactions between the MIME and IDNA encoding algorithms cannot be discounted.

- The utility of off-line reference to IDN presupposes either omnipresent IDN support in the client environment, or awareness of possible need for special configuration before, for example, an IDN can be transcribed from a printed source into the address line of a Web browser or the To line of an e-mail message. Where such information passes across boundaries between language communities, visible Punycode is likely to prove an indispensable adjunct communication format.

- On-line references to IDN can be couched in a variant of the familiar "in order to view this material you will need to download …" admonition. The target of a URL can also be an LDH equivalent to the displayed IDN, or in Punycode. Users will also need to be informed about the risks for deceptive exploitation of graphically similar characters that appear in different scripts at different code points.

- Since user exposure to Punycode may prove inevitable in a variety of contexts, preemptive support for the basic IDNA components ToASCII and ToUnicode may be worth including in tools that are rich in specialized Unicode functionality, particularly in text editors with HTML rendering capability.

Summary

The ability to provide Internet resources with localized identifiers is seemingly only a small contribution to the development of the Internet as a global multilingual content repository. However, the names of such things as Web sites appear in countless places both on-line and off. The availability of IDN may therefore be invaluable in calling attention to the manifold cultural aspects of the Internet. The clear labeling of a network facility in the language of the community by or for which that facility is maintained can provide a significant inducement to the contribution of content. In any case, it provides simple graphic means for indicating that the Internet is being ushered into a cultural universe that is far larger than the one in which it originated.

The cultural nexus of the .museum TLD, together with its controlled namespace, makes it a singularly appropriate venue for demonstrating this. The constraints that attach to the rate and

extent of IDN development directly on the second-level under a TLD are not all applicable in the administration of lower-level domains. The aggregate of subdomains under icom.museum provides a coherent platform for activity otherwise conducted in an extensive range of languages, and can therefore further serve as a useful basis for development and testing of IDN facilities in a correspondingly large range of cultural contexts.

The development of IDN has primarily been steered by the concerns of the protocol engineering community and policy maintenance agencies. This strict bimodality, however reasonable an attribute of the normative process it is, may be less appropriate in the deployment arena. In hindsight, the DNS may have been designed without full appreciation of the extent to which domain names could be seen as words. IDN must be allowed to develop in recognition of the full cultural significance of internationalizing the namespace and in a manner that fosters, rather than inhibits, the swift realization of the benefit that this can generate to content providers, to Internet users, and in maintaining the greatest possible language diversity.

Notes and references

1. The DNS was defined incrementally in a series of Requests for Comments (RFCs). A sequence of preliminary documents culminated in 1982, when Zaw-Sing Su and Jon Postel released RFC819, *The Domain Naming Convention for Internet User Applications*, http://www.rfc-editor.org/rfc/rfc819.txt.

Full protocol details were devised in a collaborative process led by Paul Mockapetris and specified in 1983 in RFC882, *Domain Names - Concepts and Facilities*, and RFC883, *Domain Names - Implementation and Specification*, http://www.rfc-editor.org/rfc/rfc883.txt, which the same author replaced in 1987 by the identically named RFC1034, http://www.rfc-editor.org/rfc/rfc1034.txt, and RFC1035 http://www.rfc-editor.org/rfc/rfc1035.txt. A number of supplementary RFCs appeared both in the interim and subsequently, describing extensions to the basic system.

The initial top-level domain names were listed in 1984 by John Postel and Joyce Reynolds in RFC920, *Domain Requirements*, http://www.rfc-editor.org/rfc/rfc920.txt. In 1994, Postel published an informational update in RFC1591, *Domain Name System Structure and Delegation*, http://www.rfc-editor.org/rfc/rfc1591.txt, which summarizes the state of the TLD namespace as initially structured and, despite being classed as informational, is the primary statement of the detail of the DNS. In light of developments that he would make public shortly thereafter, it is worth noting the author's statement that, "It is extremely unlikely that any other TLDs will be created."

2. A domain name is displayed as a sequence of labels separated by full stops, referred to as dots. Although rarely shown, there is a final dot at the end of this sequence indicating the DNS root. The form in which a domain name is transmitted on the wire does not include the dots, instead prefacing each sequence of octets that constitutes a label with an octet indicating the number of octets in that label. A single label may not contain more than 63 octets. The total number of octets in a name may not exceed 255.

The representation of a single ASCII-encoded Unicode character in an IDN label requires three octets. The number of octets needed to represent more than one encoded Unicode character in

a single label is not a linear multiple of the number needed for a single character. There is an additional overhead of four octets in every label containing one or more such characters. (The concept of ASCII-encoding is explained in elsewhere in this text.)

3. RFC920 makes reference to an unspecified number of country domains based on the two-letter English codes listed in the ISO-3166 standard. This is subject to modification by the ISO-3166 Maintenance Agency, which posts the most recent version of the list at http://www.iso.org/iso/en/prods-services/iso3166ma/02iso-3166-code-lists/list-en1.html.
It should be noted that this list contains codes for entities that are not countries, which is a source of some difficultly.

4. In 1996, Jon Postel posted an Internet Draft on *New Registries and the Delegation of International Top Level Domains*, widely referred to as "Draft-Postel". This was not intended to be cited other than as a work in progress and its latest version expired at the end of that year. (Such drafts reflect conditions that are liable to change and are therefore deliberately presented for limited periods.) It provided a detailed personal analysis of the significant shifts that had occurred since the establishment of the DNS in the way domain names were perceived. One of its conclusions was that a useful purpose would be served by the establishment of a number of new "international TLDs", proposing:

> *It is estimated that approximately thirty (30) new iTLDs allocated to approximately ten (10) new registries will be created per year. It is expected that this will continue for the next five years - unless something significant happens to change this plan. In this first year of this plan significantly more new iTLDs and registries may be chartered, perhaps up to one-hundred-fifty (150) iTLDs allocated to up to fifty (50) registries.*

5. Although a domain name is clearly terminated on its right end by the appearance of a TLD label (with utter non-ambiguity provided by the use of the final dot), the leftmost label may designate a sub-domain, a host, or both. To avoid any risk for the inclusion of illegal characters in a host name, it is common practice to apply the 'host name rule' to all labels unless there is pressing reason to do otherwise.

6. Dots are used as separators in both domain names and numeric IP addresses. If there were no restrictions on the inclusion of numbers in the former, it would be possible for a domain name to be indistinguishable from an IP address. In order to guarantee that this cannot happen, at least one label in a domain name must be non-numeric. This is ensured by disallowing numeric TLD labels, and it is common practice for registries to exclude digit-only labels on other levels.

7. There are four RFCs in the IDNA suite. In December 2002, Paul Hoffman provided the basic document in RFC3454, *Preparation of Internationalized Strings ("stringprep")*, http://www.rfc-editor.org/rfc/rfc3454.txt. This describes means for processing Unicode text so that it can be represented by displayable characters with code points below U+007F and provides a general framework for application-specific profiles.

In March 2003, Patrick Fältström, Paul Hoffman, and Adam Costello described the details of IDNA in RFC3490, *Internationalizing Domain Names in Applications (IDNA)*,

http://www.rfc-editor.org/rfc/rfc3490.txt. This required the use of a Stringprep profile, described by Paul Hoffman and Marc Blanchet in RFC3491, *Nameprep: A Stringprep Profile for Internationalized Domain Names*, http://www.rfc-editor.org/rfc/rfc3491.txt, and the encoding algorithm described by Adam Costello in RFC3492, *Punycode: A Bootstring encoding of Unicode for Internationalized Domain Names in Applications*, http://www.rfc-editor.org/rfc/rfc3492.txt. IDNA uses the Unicode version 3.2 character repertoire. The code points that may appear in Stringprep output are restricted by explicit prohibition, remapping, and normalization (NFKC).

8. The proposal made in 1996 for the creation of new generic TLDs triggered protracted and intricate debate. This ultimately led to ICANN's establishment to administer the TLD process as part of its broader responsibility for ensuring the stable operation of the DNS. Detailed information about ICANN is available at http://www.icann.org/. The IDN Guidelines are at http://www.icann.org/general/idn-guidelines-20jun03.htm with a revision in progress.

9. Several IDN libraries are available. An open source initiative is located at http://www.gnu.org/software/libidn/ where a list of additional libraries and SDKs is also provided.

10. ICOM was established in 1946 and is a non-governmental organization (NGO) maintaining formal relations with UNESCO. ICOM acts together with other heritage sector NGOs such as the International Council on Archives (ICA), the International Council on Monuments and Sites (ICOMOS), and the International Federation of Library Associations (IFLA) in charting and responding to situations where cultural property is endangered, as well as in other cross-discipline development initiatives. Details about the organization are provided at http://icom.museum/.

11. http://portal.unesco.org/ci/en/ev.php-URL_ID=13475&URL_DO=DO_TOPIC&URL_SECTION=201.html

12. http://www.remunere.net/, http://about.museum/musenic/final.report/

13. http://www.icann.org/tlds/agreements/museum/sponsorship-agmt-att1-20aug01.htm

14. Some two-letter country codes are coincidentally also dictionary words or popular abbreviations and the corresponding TLDs may be marketed for that value.

15. Two basic approaches to determining the range of available IDN characters are described in a *Briefing Paper on IDN Permissible Code Point Problems* prepared by the ICANN IDN Committee, available at http://www.icann.org/committees/idn/idn-codepoint-paper.htm. The path chosen for IDNA is clearly delineated in the RFCs. The IDN committee's underlying concern with 'homograph attack' was recognized and described long before IDNA went into design, and limiting the scope of its potential exploitation was one of the reasons for the committee having been established, in the first place, and a key focus of its deliberations.

This was the subject of intense renewed discussion at the time the present paper was being finalized. Key material presented during its course is either located at or referenced in a new IDN information area provided by ICANN at http://icann.org/topics/idn.html. The Draft Unicode Technical Report #36 *Security Considerations for the Implementation of Unicode and*

Related Technology at http://www.unicode.org/reports/tr36/tr36-2.html, provides a particularly detailed description of the basic concerns and appears likely to be used as an informal reference (pending its own finalization) in much of the work toward revising and refining IDN policies and implementations. A review of the present state of these developments may be included in the conference presentation of this text.

16. http://about.museum/policy.html

17. The references to the ICOM committees as listed in the present text display Unicode but are anchored to LDH. This ensures a functional clickpath in any client environment. The direct invocation of the Unicode form does, of course, require IDN-aware software. Alternate forms are presented side-by-side in the extended list. It should be noted that most of the references display characters that are not permitted in URLs. This further highlights the need for the implementation of IRI as described in the reference immediately below.

18. http://www.rfc-editor.org/rfc/rfc3987.txt.

The author

Cary Karp is the President and CEO of the Museum Domain Management Association, the organization responsible for the .museum TLD. He is also Director of Internet Strategy and Technology at the Swedish Museum of Natural History and serves in the same capacity for the International Council of Museums. He holds a PhD in musicology from Uppsala University where he is Associate Professor of Organology.

Dr. Karp is involved in action toward establishing further TLDs equivalent to .museum for other segments of the heritage management sector, extending collaborative action within that community onto the Internet. When IDNA was released, .museum was among the first TLDs to provide the corresponding service. Karp is active in developing the potential of IDN for designating cultural activity that does not coincide with national boundaries, with particular regard to the intangible heritage of indigenous cultures.

Cary Karp
Swedish Museum of Natural History
Box 50007
SE-10405 Stockholm
ck@nic.museum
http://about.museum/