Enabling Human Rights with Universal Acceptance: The Path to the Implementation of Internationalized Domain Names (IDN) and Email Addresses Internationalization (EAI)

Sávyo V. Morais[1], **Mark W. Datysgeld**[2]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte (IFRN)
Ceará-Mirim – RN – Brazil

[2]Governance Primer
São Paulo – SP – Brazil

savyovm@gmail.com, mark@governanceprimer.com

Abstract. *This paper seeks to understand the interactions between Human Rights and the technical aspects of linguistic representation in email systems and the DNS. Based on studies carried out concerning Universal Acceptance and the usage of non-ASCII characters on the Internet, it is proposed that this subject is of relevance to the IETF and the broader technical community that surrounds it. Going over results from recent research on the subject, a case is presented that while there have been advances in the representation of different scripts online, more can be done to ensure that every person is able to experience the network using their native language.*

1. Introduction

In little more than a half-century, the Internet expanded from a project that connected select academic institutions from the developed world into a global network that has reached most nations in the planet. This rapid progress was backed by intellectual and financial investment into the evolution of the equipment, political agreements, and technical standards that enable its operation, and many challenges are in the course of being overcome.

Concerns over how to enable unbounded written communication online emerged particularly starting from the decade of 1990, as institutions such as Unicode and its members made significant progress in facilitating the usage of different writing scripts in the Internet [Cox and Pike 2003]. While it can be said that, at present, most individuals can already produce digital content using their favored script, this cannot be said of domain names and email addresses.

The impossibility of a user having a complete online experience without the need to switch from their writing script to ASCII input was, for the longest time, seen by the broader Internet community as an undesirable but unavoidable reality. Starting from the late 2000s, but particularly from the decade of 2010, a gradual change to this thinking has been taking place, and efforts are being carried out to establish a new paradigm.

Universal Acceptance (UA) was adopted as an umbrella term for the mission of generating acceptance of all domain names and email addresses, regardless of platform or usage [Kende and Kloeden 2017]. A semi-autonomous group emerged within the Internet Corporation for Assigned Names and Numbers (ICANN) community in 2015 under the name of Universal Acceptance Steering Group[1] (UASG) to coordinate these efforts, and has ever since been producing knowledge on the theme.

Observing these growing efforts, we hold that linguistic inclusion is reflected in the Universal Declaration of Human Rights [The United Nations 1948] particularly in Article 27: the ”Right to Culture”. It is also notable that RFC 8280 discusses the internationalization of the protocols, standards, and implementations as a crucial path towards a truly global Internet [ten Oever and Cath 2017].

This paper aims to contribute towards the efforts of the Human Rights Protocols Considerations Research Group (HRPC RG) of the Internet Research Task Force (IRTF). In order to achieve this, we briefly explore the evolution of UA as a subject, as well as presenting some of the latest research performed by the authors and colleagues from the UASG community, with additional material that complements the existing data. It is our hope that, with this, comprehension of the matter at hand will be made simpler.

2. Background

The DNS is a long-standing Internet resource, and certain limitations originating from its development process still constrain Human Rights (HR) goals to this day, in spite of the several incremental improvements that were made to the system over the decades. Its first iteration consisted of a list of “host names, addresses, and attributes”, and was explicitly conceived to make use of the ASCII character encoding only, as stated in RFC 608 [Kudlick 1974].

This methodology was refined by Jon Postel and several other contributors as the network expanded, and eventually the *Domain Style Naming System* was established as a hierarchical system reliant on a single authoritative *hosts.txt* file, which essentially contained a table correlating IP addresses with domain names to make access to them easier and more reliable. Its taxonomy implicitly expected there to be three or four characters at the Top-Level Domain (TLD), such as *.edu* and *.arpa*.

While momentous for its time, the combination of the explicit use of ASCII encoding and an implicit expectation of three or four characters at the TLD level eventually became problematic as the Internet gained global reach, impacting the many people in the planet who either do not make have Latin characters as their primary script.

While the content side of the Web and email messages progressed towards better compliance with Unicode standards and a consequent support of more writing scripts, the addressing side (domain names and mail box addresses) lagged behind significantly. The almost complete lack of support for non-ASCII characters in the DNS remained a problem well into the decade of 2000, when at last IDNs had their incorporation [Klensin et al. 2006] at the second-level sanctioned by ICANN [Taylor 2011].

It was only in 2009 that ICANN’s Board approved the “IDN ccTLD Fast Track” process, after three years of consensus-building work under the Internationalized Domain Names Working Group [ICANN 2009]. By then, browser plug-ins that provided support to the usage of the Han script in all domain names already saw adoption in China and adjacent territories, and in 2006, the Chinese government was challenging ICANN’s role as

[1] Available at: `https://uasg.tech/about/`. Accessed: 05/31/2021

the sole authoritative public root, leading to increasing concerns over root fragmentation and creating pressure for action [Bray 2006].

However, the authoritative implementation of IDNs as the decade of 2010 started did not mean that those became widely known and understood, seeing as, until then, Internet-related applications and equipment had been operating under the previous assumptions about names. As important as ICANN's role in the broader Internet Governance ecosystem is, many of the decisions taken within its policy-making processes escape the perception of even those close to the technology community [Datysgeld 2018].

As a result of the "New gTLD Program", 116 applications for IDN TLDs were made, 73 of which were in the Han script, with the overall interest in the program heavily concentrated in Asia. Several ccTLDs and gTLDs have since been committed to the root and became operational, but remain affected by what has been deemed a "chicken and egg" problem, in which due to lack of demand there is little support from developers, which in turn generates little demand for the usage of the domain names, and vice versa.

After the founding of the Universal Acceptance Steering Group (UASG) in 2015, several research papers and studies were generated with the objective of attempting to address existing gaps, and are progressively finding more reach in the Internet Governance community. They have not, however, found significant readership and usage within the IETF yet.

3. Open-Source Community Survey

Intending to identify what the software landscape currently looks like in terms of compliance with UA, our study crawled all Java and Python open-source projects hosted on the GitHub platform, evaluating their usage of different libraries based on the information contained within their dependency files.

The presence of certain libraries hints at software that performs validation tasks, which may include domains and email addresses. We can further estimate the rate of success of operations involving IDNs and EAI based on the known compatibility statuses presented by UASG studies. Table 1 shows the list of Java and Python libraries that were deemed to be UA-related and their compliance status [UASG 2020].

To collect the data, we crawled the GitHub API looking for metadata from projects which contained Python Pip or Java Maven dependency files, following GitHub's recommended standards[2] for maintaining dependencies.

The crawling process returned 187,672 Python repositories and 1,185,438 Java repositories. This dataset was made up of project metadata and a list of the libraries used by each project. We also extracted the number of forks, watchers and stars of each project, which were understood to be reasonable indicators of relevance, seeing as Github does not provide a "top software" list of its own.

After analyzing the collected data, we identified that 10.54% of Java projects and 37.77% of Python projects depend on at least one UA-related library.

Table 2 shows the "readiness" classification of projects from both languages. We

[2] Available at: `https://docs.github.com/en/code-security/supply-chain-security/about-the-dependency-graph`. Accessed: 04/16/2021

Language	Library	UA Readiness
Java	icu4j	Ready
Java	libidn	Not Ready
Java	commons-validator	Not Ready
Java	validator-api	Not Ready
Java	springfox-bean-validators	Not Ready
Java	hibernate-validator	Not Ready
Java	guava	Not Ready
Java	jakarta.mail	Ready
Python	idna	Ready
Python	pyicu	Ready
Python	idna_ssl	Ready
Python	email-validator	Ready
Python	validators	Not Ready
Python	django-auth	Not Ready

Table 1. Compliance of UA-related Libraries

Language	Ready	Partially Ready	Not Ready	Ready %	Partially Ready %	Not Ready %
Java	613	318	124,004	0,49%	0,25%	99,26%
Python	69,165	1365	283	97,67%	1,93%	0,40%

Table 2. Repositories UA Readiness Rating

classified as *Ready* the projects that only depend on compliant UA-related libraries, as *Partially Ready* the ones that depend on both compliant and non-compliant libraries, and as *Not Ready* the projects that only use non-compliant libraries.

By analyzing the results described in Table 2, we identified a significant discrepancy in UA-readiness between Java and Python. While 97.74% of the Python repositories rely completely on UA-compliant libraries, only 0.69% of the Java projects are UA-ready. This suggests advanced compliance in the Python side, but signals to the need for much engagement to be performed with the Java community.

To facilitate matters, by comparing the results from Table 2 with library occurrence in Table 3, it becomes clear that most of the analyzed Java projects rely on two specific libraries which are not compliant, while most of the Python projects rely on a single compliant library. It is therefore possible to have a significant improvement in Java projects with moderate investment, by working to improve the two top libraries identified and make them compliant with UA.

Further evaluations were carried out by the team in order to assess the popularity of the projects themselves, then identifying if popular projects had better compliance rates than less notable ones.

To do this, we correlated compliance and the project relevance. To calculate Project Relevance (PR) we used Equation 1, a weighted arithmetic average which considers three relevance parameters provided by the GitHub API: Repository Forks (R_f), Repository Stars (R_s), and Repository Watchers (R_w). A multiplication factor was asso-

Language	Library	Occurrence	Occurence %
Java	hibernate-validator	62963	43.68%
Java	guava	62821	43.58%
Java	springfox-bean-validators	12501	8.67%
Java	commons-validator	4906	3.40%
Java	icu4j	886	0.61%
Java	jakarta.mail	49	0.03%
Java	libidn	29	0.02%
Java	validator-api	2	0.00%
Python	idna	70789	95.81%
Python	validators	1660	2.25%
Python	email-validator	1177	1.59%
Python	pyicu	243	0.33%
Python	idna_ssl	10	0.01%
Python	django-auth	5	0.01%

Table 3. Occurrence of UA-related libraries

	Low Relevance			Medium Relevance			High relevance		
	Ready	Part. Ready	Not Ready	Ready	Part. Ready	Not Ready	Ready	Part. Ready	Not Ready
Java	0,43%	0,23%	99,34%	0,68%	0,35%	98,97%	2,19%	1,46%	96,35%
Python	97,75%	1,95%	0,30%	97,54%	1,88%	0,58%	95,21%	1,44%	3,35%

Table 4. Repositories UA Readiness Rating – Per Relevance

ciated to each parameter to assure weight balancing between them. The parameter multiplication factor is represented by α_X, where $_X$ is the parameter, and its value is calculated by Equation 2.

$$PR = \frac{R_f * \alpha_F + R_s * \alpha_S + R_w * \alpha_W}{\alpha_F + \alpha_S + \alpha_W} \tag{1}$$

$$\alpha_X = \frac{max(avg(F), avg(S), avg(W))}{avg(X)} \tag{2}$$

After calculating their relevance value, we categorized the repositories as follows: **Low Relevance** for repositories with $PR == 0$; **Medium Relevance** for repositories with $0 < PR <= 300$; **High Relevance** for repositories with $PR > 300$. The PR threshold values for the groups were defined arbitrarily based on histogram analysis. We then calculated the readiness of each group.

The results demonstrated a similar UA-readiness distribution regardless of relevance group, as shown in Table 4. Therefore, there is no objective advantage or disadvantage to investing in one group or another, meaning that for maximum impact to be achieved, High Relevance projects should be targeted for remediation in the near future.

4. Final considerations

Awareness in relation to Universal Acceptance has increased significantly in the past few years within the ICANN community, as well as in other groups of interest from the broader Internet Governance community. During ICANN 70, which took place in March of 2021, two sessions were carried out dedicated exclusively to the subject, as has been the case in recent meetings, with anywhere between two to four sessions on the theme.

This is not enough. It is important that the Internet Governance community as a whole is informed of these developments and that knowledge does not remain siloed within ICANN. The effort needed to accomplish the required changes to make all components UA-ready demands cooperation between all stakeholders.

It is our hope that the brief glimpse provided here is one of many opportunities in which knowledge on this subject is ported over to the IETF, and concerted efforts and productive discussions can be carried out to advance linguistic diversity and HR goals on the Internet. More studies are already in the UASG's pipeline, and can be followed on the project's website.

References

Bray, H. (2006). China sets up system for Internet domains. https://nytimes.com/2006/03/01/technology/china-sets-up-system-for-internet-domains.html. Accessed: 04/16/2021.

Cox, R. and Pike, R. (2003). UTF-8 History. https://www.cl.cam.ac.uk/~mgk25/ucs/utf-8-history.txt. Accessed: 06/03/2021.

Datysgeld, M. (2018). Understanding the Role of States in Global Internet Governance: ICANN and the Question of Legitimacy. In *Proceedings of the Annual Symposium of the GigaNet: Global Internet Governance Academic Network*. SSRN.

ICANN (2009). Adopted Board Resolutions: Seoul. https://icann.org/resources/board-material/resolutions-2009-10-30-en. Accessed: 04/16/2021.

Kende, M. and Kloeden, A. (2017). Unleasing the power if all domains: The social, cultural and economic benefits of universal acceptance. Technical report, Analysys Mason Limited.

Klensin, J. C., Fältström, P., Karp, C., and IAB (2006). Review and Recommendations for Internationalized Domain Names (IDNs). RFC 4690.

Kudlick, M. (1974). Host names on-line. RFC 608.

Taylor, E. (2011). *Internationalised domain names: state of play*. EURid, Belgium.

ten Oever, N. and Cath, C. (2017). Research into Human Rights Protocol Considerations. RFC 8280.

The United Nations (1948). *Universal Declaration of Human Rights*.

UASG (2020). Universal Acceptance Compliance of Some Programming Language Libraries and Frameworks. Technical Report 18a, Universal Acceptance Steering Group (UASG).