The final authenticated version is available online at https://doi.org/10.1007/978-3-030-15986-3_12

Funny Accents: Exploring Genuine Interest in Internationalized Domain Names

Victor Le Pochat[0000−0003−2297−8328], Tom Van Goethem, and Wouter Joosen

imec-DistriNet, KU Leuven, 3001 Leuven, Belgium
{firstname.lastname}@cs.kuleuven.be

Abstract. International Domain Names (IDNs) were introduced to support non-ASCII characters in domain names. In this paper, we explore IDNs that hold *genuine interest*, i.e. that owners of brands with diacritical marks may want to register and use. We generate 15 276 candidate IDNs from the page titles of popular domains, and see that 43% are readily available for registration, allowing for spoofing or phishing attacks. Meanwhile, 9% are not allowed by the respective registry to be registered, preventing brand owners from owning the IDN. Based on WHOIS records, DNS records and a web crawl, we estimate that at least 50% of the 3 189 registered IDNs have the same owner as the original domain, but that 35% are owned by a different entity, mainly domain squatters; malicious activity was not observed. Finally, we see that application behavior toward these IDNs remains inconsistent, hindering user experience and therefore widespread uptake of IDNs, and even uncover a phishing vulnerability in iOS Mail.

Keywords: Internationalized Domain Names · Phishing · Domain squatting · Homograph attack.

1 Introduction

The Internet has become a global phenomenon, with more than half of the world's households being estimated to have Internet access [2]. The English language and Latin alphabet remain dominant, but multilingual content is enjoying increased popularity [19, 59]. However, one crucial part of the Internet, the Domain Name System (DNS), has historically been limited to ASCII characters [5, 27, 46].

Internationalized Domain Names (IDNs) [20, 35] have been introduced to address this problem, and domain names can now contain (Unicode) characters from various languages and scripts. IDNs allow end users to refer to websites in their native language, and have helped to increase linguistic diversity, with a strong correlation between a website's language and the script of its IDN [19].

Acceptance of IDNs relies on support by web applications, and while this has been improving, significant gaps that present a barrier to user recognition and adoption remain [19]. Moreover, IDNs have seen abuse, with malicious actors registering domains that use visually similar characters to impersonate popular domains for phishing attacks [21, 28, 41]. This further complicates how browsers choose between displaying IDNs and protecting end users [1, 44].

In this paper, we explore (ab)use of IDNs for over 15 000 popular brands and phrases that contain non-ASCII characters (e.g. "Nestlé"), obtained through the presence of their ASCII equivalent in a set of popular domains (`nestle.com`). For these, we define IDNs that hold *genuine interest* (`nestlé.com`): these IDNs can enhance user experience as they are easier and more natural to read and correctly understand, and both end users and brand owners may therefore prefer to use them. Moreover, country-specific keyboard layouts often feature dedicated keys for characters with accents, making typing them no more difficult than non-accented letters. We study whether owners of popular domains where an IDN with genuine interest exists have made the effort to register and use it.

However, these IDNs can also attract malicious activity. While previous work studied abuse of IDNs resembling very popular brands [41], these brands generally do not feature accents, meaning that users are less prone to use or trust the IDNs, and brand owners are not inclined to own them except for defensive purposes. In contrast, as our IDNs with genuine interest appear 'valid' to end users, it becomes even more difficult to distinguish a legitimate website from an attempt at phishing, and the domains are therefore more valuable to malicious actors. This also enables attacks akin to typosquatting [16], as users may type the (non-)accented version of a domain, even though this may host a different website. We determine whether these IDNs are still open for or already see abuse.

In summary, we make the following contributions: (1) we generate 15 276 candidate IDNs with genuine interest as derived from the page titles of popular domains; (2) we see that 43% can still easily be registered, e.g. for domain squatting or abuse by malicious parties; (3) we estimate at least 50% of the IDNs to share ownership with the original domain, but 35% to have different owners, mostly domain squatters; (4) we see that browsers and email clients display IDNs inconsistently: our survey even leads us to discover a vulnerability in iOS Mail that enables phishing for domains with `ß`.

2 Background and related work

Internationalized Domain Names Through the Domain Name System (DNS), user-friendly domain names are translated into IP addresses. Domain names represent a hierarchy, with the registries managing the top-level domains (e.g. `.com`) usually delegating the public offering of second-level domains (e.g. `example.com`) to registrars. Originally, the *LDH* convention restricted domain names to ASCII *l*etters, *d*igits and *h*yphens [5,27,46]. However, languages like French and German use Latin characters with diacritics, and e.g. Arabic and Chinese use different character sets altogether. To provide a universal character encoding of these writing systems, the Unicode Standard [65] was developed.

To support domain names with Unicode labels, IETF developed the Internationalized Domain Names in Applications (IDNA2003) protocol in 2003 [20]. To maintain compatibility with existing protocols and systems, this protocol uses the Punycode algorithm [10] to convert Unicode labels ("U-label") to an ASCII Compatible Encoding (ACE) label starting with `xn--` and containing only

ASCII characters ("A-label"). In 2010, the standard was revised (IDNA2008) [35], mainly to add support for newer versions of the Unicode Standard.

Homograph attacks Homographs are strings that contain homoglyphs or visually resembling characters, and can be used to trick users into thinking that they are visiting one domain while actually browsing another, opening up opportunities for web spoofing or phishing [14, 28]. While certain ASCII characters (e.g. lower case `l` and upper case `I`) already allowed for confusion, the introduction of IDNs gave rise to a whole new set of potential homographs, using either diacritics or resembling characters from other scripts. Evaluations over time of browser and email client behavior regarding IDNs have found that browsers have implemented countermeasures in response to vulnerabilities to homograph attacks, but that they are not (yet) fully effective [24–26, 41, 45, 71].

Previous studies have shown IDNs confusable with popular domains to exist on a modest scale and for relatively benign purposes such as parking [21, 28]. In 2018, Liu et al. [41] detected 1 516 out of 1.4 million registered IDNs to exploit homographs for targeting domains in Alexa's top 1 000. Only 4.82% belonged to the same owner as the original domain. Moreover, they generated 42 434 additional IDNs with sufficient visual similarity that are still unregistered. Tian et al. [66] searched for phishing sites that impersonate a set of 702 popular brands both in content and in domain, a.o. through homograph domains. Several industry reports have addressed homograph attacks in the wild, seeing circumvention of spam filters [70], phishing, malware and botnet abuse [38] and popular as well as financial websites being main targets [56].

Domain squatting Domain names can be exploited for deceiving end users: involuntary errors redirect traffic to unintended destinations [3, 15, 16, 50, 63, 67, 69], while credible domain names may create the perception of dealing with a legitimate party [34, 43, 48]. Spaulding et al. [61] reviewed techniques to generate, abuse and counteract deceptive domains. Liu et al. [41] found 1 497 IDNs that combine domains from Alexa's top 1 000 with keywords containing non-ASCII characters. They also mention a type of abuse where the IDN is the translation of a brand name to another language, but do not conduct any experiments.

3 Methods

3.1 Generating candidate domains

In order to obtain IDNs with genuine interest, we start from a list of popular domains. While the Alexa top million ranking is commonly used, Scheitle et al. [55] and Le Pochat et al. [39] have shown that it has become very volatile and disagrees with other rankings, while the latter proved that manipulation by malicious actors requires very low effort. Therefore, we use the Tranco list[1] proposed by Le Pochat et al. [39], a list of one million domains generated by

[1] https://tranco-list.eu/list/RQ4M/1000000

Table 1. Candidate IDNs are generated by searching relevant substitutions within a domain name using its root page title.

Original domain	Root page title	Converted to lowercase, punctuation removed	Diacritics removed/ substitutions applied	Derived IDN
example.com	Example domain	example domain	example domain	*No IDN*
nestle.com	Home | Nestlé Global	home nestlé global	home nestle global	nestlé.com
uni-koeln.de	Universität zu Köln	universität zu köln	universitat zu koln universitaet zu koeln	– uni-köln.de

combining four rankings over 30 days (here 30 July to 28 August 2018), in order to require prolonged popularity from multiple vantage points.

We check for each domain whether it corresponds to a string that contains diacritical marks, i.e. where there could be genuine interest in adopting a variant IDN. For this purpose, we look for plausible substitutions with accented words in the title of its root page. To collect these title strings, we use a distributed crawler setup of 4 machines with 4 CPU cores and 8 GB RAM, using Ubuntu 16.04 with Chromium version 66.0.3359.181 in headless mode.

We then convert this title to lowercase and remove punctuation, after which two strings are generated: either diacritical marks are simply removed, or language-specific substitutions are applied (as listed in Appendix A). The latter covers the common practice in for example German to use replacements such as `ae` for `ä`. We then compare these converted (ASCII) strings with the domain name: we favor the case where the full domain is found, but also consider cases where single words are shared. Finally, if such cases are found, we retrieve the corresponding accented form from the original title and apply this substitution to the original domain name, resulting in the candidate IDN. Table 1 illustrates our approach.

3.2 Retrieving domain-related data

To understand if and how these IDNs are used, we collect the following data:

DNS records To check whether candidate IDNs exist in the DNS (i.e. are registered) and how they are configured, we request `A`, `MX`, `NS` and `SOA` records for both the original and candidate domain. If all records return an `NXDOMAIN` response, we assume the domain to be unregistered. Otherwise, we verify whether the nameserver is properly set up (no `SERVFAIL`) and if there are `A` records (suggesting a reachable website) or only other records (suggesting another purpose).

Domain eligibility A TLD registry is free to support IDNs or not, and if they do, they may only allow a specific set of characters. For country code TLDs this set usually consists of the characters in languages spoken in that TLD's country, which can help in avoiding homograph attacks by prohibiting confusable characters that would normally not be used in those languages.

ICANN's IDN guidelines [29] require registries to publish "Label Generation Rulesets" (LGR), i.e. lists with permitted Unicode code points, in IANA's Repository for IDN Practices [30]. However, as of this publication, only six TLDs had

published these machine readable LGRs. For 626 other TLDs, the repository contains simple text files that list the code points. Where possible, we parse these files and generate the corresponding LGRs with ICANN's LGR Toolset [31]. For the remaining TLDs, no information is available from the repository. We manually search the IDN policy and generate an LGR for 30 additional TLDs. Finally, we validate our candidate domains against these LGRs with the LGR Toolset to determine whether they are allowed by their respective registries.

Domain availability To determine whether unregistered domains can be readily bought through a popular registrar, we query GoDaddy's API [22] for their availability. This data complements the eligibility data, as further restrictions may apply for certain TLDs (e.g. being based in that TLD's country): in this case the API returns an error indicating that the TLD is unsupported, otherwise the API returns whether the domain is (un)available.

WHOIS records To obtain ownership information for the domains in our data set, we retrieve and parse their WHOIS records with the Ruby Whois library [7]. However, WHOIS data has several limitations, especially for bulk and automated processing. The format of WHOIS data varies widely between providers (which can be registries or registrars); it may be human-readable, but both parser-based and statistical methods cannot retrieve all information flawlessly [42]. Moreover, rate limits prevent bulk data collection.

Even if data can be adequately obtained, it may not be of high quality. Registrant details can contain private contact information, so privacy concerns and malicious intent have spurred a number of privacy and proxy services, whose details replace those of the real owner [9]. The European General Data Protection Regulation (GDPR) has also cast doubt on whether such data can still be released [32], with e.g. the `.de` registry already withholding any personal details [13]. Finally, WHOIS data may be outdated, e.g. not reflecting company name changes, or the same registrant may use different data across domains.

Web pages To determine what content the accented and non-accented domains serve, we visit the root page for each domain pair where the IDN has a valid `A` record. By limiting our crawl to one page, we minimize the impact on the servers hosting the websites. As with our title crawl, we use a real browser to capture the request and response headers, the redirection path and final URL of the response, TLS certificate data, the HTML source and a screenshot.

To classify domains, we first compute a perceptual hash of the screenshot based on the discrete cosine transform [37]. As visually similar images have similar hash values, we cluster their pairwise Hamming distances using DBSCAN [18] to find groups of websites with (nearly) the same content, which we then manually label. We also compare the hashes of the original domain and its IDN to detect equal but non-redirecting domains. Finally, for domains that were not classified using their hash, we check for the presence of certain keywords (e.g. 'parking') in the HTML source, or else decide that we cannot classify the domain.

Blacklists To detect whether our candidate IDNs exhibit malicious behavior, we match them and the domains they redirect to against the current blacklists provided by Google Safe Browsing [23] (malware and phishing), PhishTank [53]

Table 2. Summary of the registration properties of our candidate IDNs.

Candidates	15 276	(100.0%)		
Unregistered	12 087	(79.1%)	Readily available	6 608 (54.7%)
			Unavailable/Additional restrictions	4 116 (34.1%)
			Non-compliant with TLD policy	1 363 (11.3%)
Registered	3 189	(20.9%)		

(phishing), Spamhaus DBL [60] (spam), SURBL [62] (spam, phishing, malware and cracking) and VirusTotal [8] (malware).

3.3 Limitations

We restrict our search to IDNs with variations on characters of the Latin alphabet. Our exploration could be broadened to popular domains that are a romanized (converted to Latin alphabet) version of brands or phrases in another character set. However, a script often has multiple romanization standards that may be language-dependent [64]: for example, `Яндекс` (Yandex) can be romanized to `Iandeks`, `Jandeks` or `Yandeks`. We therefore ignore other character sets to avoid false positives and negatives caused by these differing systems.

Our approach to select candidate IDNs is conservative: our requirement that whole words from the title and domain match, may mean that we miss some candidate IDNs, e.g. if the domain is an abbreviation of words in the title. However, through this approach we limit erroneous candidate IDNs, which we estimate would more likely be either unregistered or maliciously used, as no one would have a genuine interest in owning the domain.

4 Results

In this section, we determine whether IDNs with genuine interest share ownership with the popular domain they are based on, and for what purpose they are used. Through a crawl conducted between 30 August and 28 September 2018, we were able to retrieve a non-empty title from the root page of 849 341 out of 1 million domains (website rankings are known to contain unreachable domains [39]). Using the process described in Section 3.1, we generated 15 276 candidate IDNs.

4.1 Registration and ownership

Table 2 lists whether our candidate IDNs with genuine interest are still available for registration. Of the 79.1% unregistered IDNs, 11.3% do not comply with their respective TLD's LGR policy, meaning that an owner of a popular domain cannot register the corresponding IDN and loses out on the user experience benefits. Through the GoDaddy API, we find that 43.3% of all candidate IDNs are readily available; 26.9% are unavailable for registration, because the registry either

Table 3. Summary of the classification of the registered IDNs with genuine interest.

(a) Domain ownership.

Same owner	1 595 (50.0%)
Same configuration	289 (9.1%)
Different owner	1 102 (34.6%)
Insufficient data	203 (6.4%)

(b) DNS records.

`A` same	704 (22.1%)
`NS` or `SOA` same	736 (23.1%)
`NS` and `SOA` different	624 (19.6%)
All records different	838 (26.3%)
Other/no data	287 (9.0%)

(c) WHOIS records.

Same contact	319 (10.0%)
Same nameserver	378 (11.9%)
Other	923 (28.9%)
No data	1 569 (49.2%)

(d) Website availability.

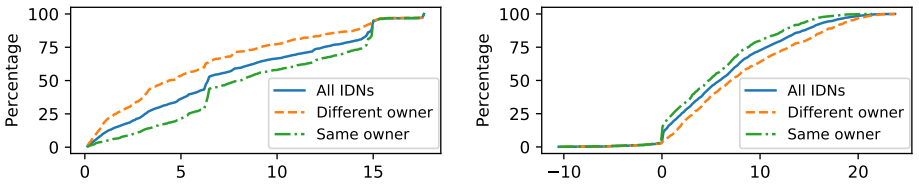
No A record	455 (14.3%)
HTTP status 200	2 466 (77.3%)
Other HTTP status	160 (5.0%)
Not reachable	108 (3.4%)

(e) Website content.

Redirect to original	1 215 (49.3%)
Identical content	112 (4.5%)
Parked/for sale	751 (30.5%)
Empty/default	132 5.4%)
Unknown	256 (10.4%)

(f) TLS setup.

Same certificate	479 (22.1%)
Different certificate	1 687 (77.9%)
Secure	171 (7.9%)
Insecure	964 (44.5%)
No connection	1 031 (47.6%)



(a) Years between the creation of the IDN and October 15, 2018, for 3 IDN types.

(b) Difference in years between the creation of the original domain and the IDN.

Fig. 1. Cumulative distribution functions for the creation dates of registered IDNs.

blocks visually similar registrations or applies further restrictions to registrants, which could also increase the burden for a malicious registration.

For the 20.9% registered domains, we compare the DNS (Table 3b) and WHOIS (Table 3c) records and web crawl data (Tables 3e and 3f) to estimate whether the original domain and its IDN have the same owner (summarized in Table 3a). For 50.0%, we believe both domains to have the same owner: they have overlapping WHOIS contact data, have the same `A` record, serve the same web content and/or present a TLS certificate for the same domains. For an additional 9.1%, shared nameservers or `SOA` records also allow us to reasonably assume shared ownership. For 34.6%, we believe both domains to have a different owner: either their `NS` and `SOA` records are both different, or the domain is parked or for sale. Brand owners would be unlikely to use the latter for monetizing their IDN, as they could better serve the actual website the visitor is looking for, and the domain would not be displaying content from a third party.

Figure 1 shows the distribution of creation dates of the IDNs. Brand owners tend to have registered their IDNs earlier than average, while domain squatters registered them later (Figure 1a). The majority of IDNs was registered after the original domain, although 3.7% of IDNs were registered earlier (Figure 1b).

In our data set, we can see examples of companies that do or do not cover IDNs when protecting their brand on the Internet. Nestlé, L'Oréal, Mömax and Citroën own several candidate IDNs, usually redirecting to the original domain, but still see some owned by third parties for parking. We also see 40 IDNs bought by brand protectors such as CSC, Nameshield and SafeBrands for their clients. However, the lack of support for certain characters hinders some companies in owning IDNs with genuine interest: e.g. the `S̃` character in Škoda sees little support by TLD registries, causing relatively low IDN ownership.

4.2 Usage

Table 3d lists whether the IDNs host a website: 14.3% of registered IDNs have no configured `A` record, suggesting proactive registration without the intention to use the IDN. Table 3e lists what content the domains that returned HTTP status code 200 serve, with 53.8% displaying the same content as the original domain, meaning that they are very likely owned and operated by the same entity. 112 IDNs are even treated equally by not redirecting to the original; however, none of the original domains redirect to the IDN. 30.5% are parked/for sale, while 5.4% show an empty/default page (e.g. unconfigured server).

Manual inspection of the domains that could not be classified shows that these largely fall into two categories. The first consists of websites that are completely different to the original domain, owned by another entity. This can leverage the popularity of the original domain, and is an opportunity to own domains with desirable phrases, but also exposes end users to confusion and potential misdirection. The second has the IDN showing slightly different or older versions of the original domain. This indicates that they both belong to the same owner and that there was an intention to use the IDN, but that it was forgotten when the original domain was reconfigured and now points to an outdated website.

4.3 Security

Incidence on blacklists is very low: none of our candidate IDNs, nor the domains they redirect to appear on the Google Safe Browsing, PhishTank, Spamhaus or SURBL blacklists. VirusTotal reports malware detections on 5 domains, but only by at most 3 out of 67 engines; these detections appear to be based on outdated information. However, Tian et al. [66] have found that over 90% of phishing sites served through squatting domains could evade blacklisting, meaning that phishing may already be much more prevalent on our candidate IDNs. Finally, parked domains are known to only sometimes redirect to malicious content [68]: we manually saw instances of such intermittent redirects to blacklisted sites for several IDNs.

Through inspection of the redirection paths, we found no proof of affiliate abuse on IDNs (sending users to the intended domain, but adding an affiliate ID to earn a sales commission), as has been seen for several domain squatting techniques [47]. We manually found examples of other, questionable behavior: `pokémongo.com` offers a "cheat code" in an online survey scam [33], and has

a cryptocurrency miner [17, 54]; jmonáe.com redirects to the original domain through an ad-based URL shortener [49]; and www.preußische-allgemeine.de includes the site of a competing newspaper in a frame (Figure 2).

From the WHOIS records, we find 81 domains to use a privacy/proxy service; while abusive domains tend to use such services [9], using them does not reliably demonstrate malicious intent [36]. Moreover, privacy concerns as well as the GDPR make that some registries and registrars hide private information by default, reducing the need to procure a privacy/proxy service.

As the web is rapidly adopting HTTPS, IDNs will also need a correct TLS setup for users to reach them without trouble. However, for the 2 166 reachable IDNs in our TLS crawl, Table 3f shows that only 7.9% are securely configured and would not cause a browser warning. The other domains either have an insecure setup (mostly because the presented certificate does not cover the IDN) or do not allow a TLS connection to be established.

For the domains with shared ownership, 60.2% are insecure or don't allow a TLS connection even though the original domain is securely configured. For 360 (26.9%) IDNs, the presented certificate is valid only for the original domain, suggesting that the domain owner has set up the original domain and the IDN identically, but has forgotten to obtain a certificate that is also valid for the IDN.

5 User agent behavior

Throughout the DNS protocol, the A-label (Punycode) of an IDN is used to maintain backward compatibility. However, developers of user interfaces may elect to display the U-label (Unicode) to provide the best user experience, as the A-label is less readable (e.g. köln.de becomes xn--kln-sna.de). In this section, we discuss the behavior of user agents regarding IDNs with diacritical marks from the Latin script, where the lack of homoglyphs makes abuse more difficult to prevent. We also uncover two edge cases that have an impact both on the value of IDNs to brand owners and on the vulnerability to IDN abuse.

Table 4 shows that popular web browsers and email clients vary widely in whether they show the A- or U-label when visiting a website or receiving email. The Gmail app on Android is a particular case, as it shows either the U-label or the A-label when email is received on a Gmail or IMAP account respectively.

Browsers based on Chromium, such as Chrome and several Android browsers, implement a special policy toward IDNs resembling very popular domains: the A-label is shown when the domain with diacritics removed appears on a hardcoded list based on Alexa's top 10 000 [1]. This policy affects 125 candidate IDNs, of which 74 are registered with 21 having the same owner: these cannot choose to prefer the IDN without affecting user experience. 2 domains already do not redirect, causing the display of the A-label. The seemingly arbitrary cut-off [58], manual addition of domains and lack of updates [57] suggest that this heuristic solution using a hardcoded list still leaves room for successful spoofing attacks.

Another edge case was introduced during the revision of the IDNA standard. Four characters (so-called "deviations") are valid in both versions, but are inter-

Table 4. Browser and email client behavior regarding IDNs with diacritical marks. For the top 10 000 `pokémon.com` was tested, for the other sites `böll.de`, and for "deviation" characters `straße.de`. 'A' denotes the display of the A-label, 'U' of the U-label. Appendix B lists the browser and email client versions used in our survey.

(a) Web browsers

		10K	other	ß/ss
Desktop	Chrome	A	U	ss
	Firefox	U	U	ß
	Safari	U	U	ß
	Opera	A	U	ss
	Internet Explorer	A	A	ss
	Edge	A	A	ss
Mobile	Chrome	A	U	ss
	Safari	U	U	ß
	Firefox	U	U	ß
	UC Browser	A	A	ß
	Samsung Internet	A	U	ss
	Opera	A	U	ss
	Microsoft Edge	A	U	ss

(b) Email clients

		10K	other	ß/ss receive	ß/ss send
Desktop	Outlook	U	U	empty	ss
	macOS Mail	A	A	A (ß)	ss
	Thunderbird	A	A	A (ß)	ß
Mobile	Gmail	U	U	U (ß)	ss
	Gmail (IMAP)	A	A	A (ß)	fails
	Outlook	A	A	A (ß)	fails
	iOS Mail <12.1.1	U	U	ss	ß
	iOS Mail ≥12.1.1			U (ß)	
Webmail	Gmail	U	U	A (ß)	ss
	Yahoo	A	A	A (ß)	fails
	Yandex	U	U	A (ß)	ss
	Outlook	A	A	A (ß)	ss
	RoundCube	U	U	A (ß)	ß

preted differently [12]: for example, the German ß is supported as-is in IDNA2008 but converted to ss in IDNA2003[2]. This results in two different domains, but the visited domain depends on which version of the standard a browser implements.

This does not only affect user experience, i.e. when links on web pages or outside the browser (e.g. in emails) point to different resources, but also has security implications. The ß domain may host a spoofing or phishing site replicating that of the ss domain [12]. Moreover, resources included from an ß domain could originate from another domain in different browsers, allowing to insert malicious content. Requiring the same owner for both domains will prevent such attacks, although errors due to misconfigured websites may persist. However, for example even the German .de registry does not currently enforce this for ß and ss.

Unfortunately, Table 4a shows that major browsers do not agree on which IDNA standard to implement, causing them to direct users to different websites as shown in Figure 2. An ß character occurs in 55 candidate IDNs, of which 26 are registered, including several bank websites. 9 domains do not belong to the same owner: the ß domain is then almost unreachable from Chromium-based and Microsoft browsers (users would have to type or follow a link to the already converted A-label), and there is potential for phishing or spoofing attacks.

Email clients also handle domains with ß differently, even between receiving and sending (Table 4b). On Outlook, the sender field remains empty. More wor-

[2] The other deviations are the Greek ς, converted to σ in IDNA2003, and the zero width non-joiner and joiner, both deleted by the IDNA2003 Punycode algorithm.

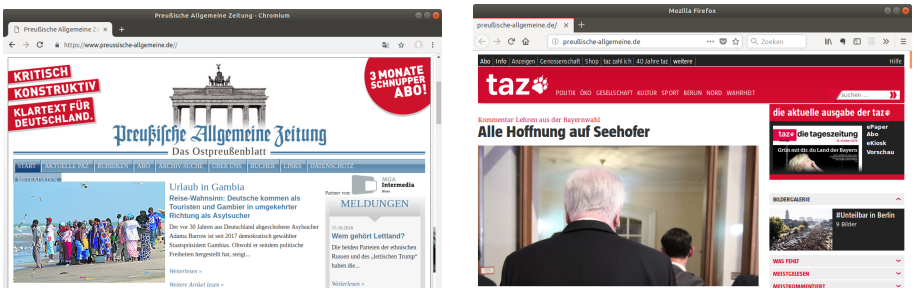


Fig. 2. Visiting `preußische-allgemeine.de` in Chrome and Firefox leads to different sites: `preussische-allgemeine.de` and `xn--preuische-allgemeine-ewb.de`.

ringly, we found that iOS Mail displayed an email received from an ß domain (e.g. `user@straße.de`) as coming from the domain with `ss` (`user@strasse.de`). This vulnerability enables phishing attacks by the owner of the ß domain; moreover, checks such as SPF will succeed as they are carried out by the mail exchangers and not the client. A reply will also be sent to the ß domain, potentially leaking sensitive information to a third party. We disclosed this vulnerability to Apple, and it was fixed in iOS 12.1.1 [4], which now displays the correct U-label.

6 Discussion

As registries are ultimately responsible for managing which domains can be registered and who can own them, they are in a prime position to combat IDN-related abuse. The most recent version of ICANN's IDN implementation guidelines [29] calls for registries to prohibit registrations of domain name variants with accented or homoglyph characters, or limit them to the same owner [40]. While certain registries implement these measures [6, 11, 51, 52], other registries that support IDNs usually either only apply such policies to homograph domains but not domains with diacritics, or do not impose any restriction at all, allowing malicious actors or domain squatters to register the IDNs with genuine interest.

On the client side, browsers and email clients represent the most visible and widespread use of IDNs. However, we have shown that they do not yet universally support the display of IDNs in Unicode, degrading the user experience. Moreover, measures put in place by browser vendors to prevent homograph attacks have been shown to be insufficient on multiple occasions [21, 41, 71]; we have done the same for a popular email client. Mozilla has expressed the opinion that registries are responsible for preventing IDN abuse, and that browser restrictions risk degrading the usefulness of IDNs [44]. Indeed, the manually developed and heuristic-based defenses cannot be expected to comprehensively solve this issue. Other protection mechanisms such as TLS and SPF also cannot prevent these attacks, as e.g. certificates can legitimately be acquired for the malicious IDN.

Owners of popular brands and domains can register the IDN with genuine interest, either as a real replacement or supplementary domain, or to proactively

stop others from abusing it. However, while this may be enough to combat (more dangerous) abuse of the 'valid' IDN with genuine interest, registering all other variant domains with homoglyphs, diacritics, and potential typos quickly becomes infeasible in terms of cost and coverage. Shared ownership of IDNs with genuine interest is already much more common than of other homograph IDNs (over 50% vs. almost 5% [41]). However, it is still concerning that at least 35% allow third parties to take hold of the valuable IDNs with genuine interest.

An unfortunate outcome of the issues surrounding IDNs would be to discourage the adoption of IDNs and to recommend that users distrust them. IDNs enable anyone to use the Internet in their native language, providing them a great benefit in user experience. IDNs also allow companies to create a better integration of brands with their Internet presence, e.g. combining a logo with a TLD in marketing material, providing additional economic value.

7 Conclusion

We have introduced the concept of Internationalized Domain Names for which there is *genuine interest*: domains that represent popular brands or phrases with diacritical marks. By comparing the page titles and domain names for 849 341 websites, we generated 15 276 such IDNs. We find 43% of them to be available for registration without restrictions, leaving the opportunity for a third party to exploit the IDN. For the 3 189 registered domains, we see that ownership is split: at least half have the same owner and content as the original domain, but at least a third belongs to another entity, usually domain squatters who have put the domain up for sale. The IDNs are not known to exhibit malicious activity, although cases of questionable behavior can be found. From insecure TLS setups and IDNs showing old versions of the original domain, we can see that brand owners who registered IDNs tend to 'forget' configuring them properly. Finally, we find applications to treat IDNs with diacritical marks inconsistently, displaying Unicode or a less readable alternative depending on resemblance to a popular domain or on the implemented version of the IDNA standard. We even found a phishing vulnerability on iOS Mail, where the actual sender domain differs from the one displayed. While brand owners have already somewhat found their way to IDNs with genuine interest, and while registries and browser vendors start to deploy tools to prevent IDN abuse, support for IDNs remains challenging, which unfortunately does not encourage their uptake in the near future.

Acknowlegdments We would like to thank our shepherd Ignacio Castro for his valuable feedback, and Gertjan Franken and Katrien Janssens for their help in the user agent survey. This research is partially funded by the Research Fund KU Leuven. Victor Le Pochat holds a PhD Fellowship of the Research Foundation - Flanders (FWO).

A Common character substitutions

Original	ä	ö	ü	ß	æ	ø	å	œ	þ
Substitution	ae	oe	ue	ss	ae	oe	aa	oe	th

B Tested user agent versions

	Client	Version	Operating system
Browser desktop	Google Chrome	69.0.3497.100	Ubuntu Linux 18.04.1
	Firefox	62.0	Ubuntu Linux 18.04.1
	Safari	12.0.1 (13606.2.100)	macOS 10.13.6 (17G65)
	Opera	55.0.2994.61	Ubuntu Linux 18.04.1
	Internet Explorer	11.0.9600.18894	Windows 8.1
	Microsoft Edge	42.17134.1.0	Windows 10 17.17134
Browser mobile	Google Chrome	69.0.3497.100	Android 7.0.0
	Safari	–	iOS 12.0 (16A366)
	Firefox	62.0.2	Android 7.0.0
	UC Browser	12.9.3.1144	Android 7.0.0
	Samsung Internet	7.4.00.70	Android 7.0.0
	Opera	47.3.2249.130976	Android 7.0.0
	Microsoft Edge	42.0.0.2529	Android 7.0.0
Email desktop	Outlook 2016	16.0.4738.1000	Windows 10 17.17134
	macOS Mail	11.5 (3445.9.1)	macOS 10.13.6 (17G65)
	Thunderbird	52.9.1	Ubuntu Linux 18.04.1
Email mobile	Gmail	8.9.9.213351932	Android 7.0.0
	Outlook	2.2.219	Android 7.0.0
	iOS Mail	–	iOS 12.0 (16A366) iOS 12.1.2 (16C104)
Webmail	Gmail	–	–
	Yahoo	–	–
	Yandex	–	–
	Outlook	–	–
	RoundCube	1.2.9	–

References

1. IDN in Google Chrome, https://dev.chromium.org/developers/design-documents/idn-in-google-chrome
2. Measuring the information society report 2017 - volume 1. Tech. rep., International Telecommunication Union (2017), https://www.itu.int/en/ITU-D/Statistics/Documents/publications/misr2017/MISR2017_Volume1.pdf
3. Agten, P., Joosen, W., Piessens, F., Nikiforakis, N.: Seven months' worth of mistakes: A longitudinal study of typosquatting abuse. In: 22nd Annual Network and Distributed System Security Symposium. Internet Society (2015). https://doi.org/10.14722/ndss.2015.23058
4. Apple Inc.: About the security content of iOS 12.1.1 (Dec 2018), https://support.apple.com/en-us/HT209340
5. Braden, R.: Requirements for internet hosts - application and support. RFC 1123 (Oct 1989)
6. Canadian Internet Registration Authority: Domains with French accented characters (Jan 2018), https://cira.ca/register-your-ca/domains-french-accented-characters
7. Carletti, S.: Ruby Whois, https://whoisrb.org/
8. Chronicle: VirusTotal, https://www.virustotal.com
9. Clayton, R., Mansfield, T.: A study of Whois privacy and proxy service abuse. In: 13th Annual Workshop on the Economics of Information Security (2014)
10. Costello, A.: Punycode: A Bootstring encoding of Unicode for internationalized domain names in applications (IDNA). RFC 3492 (Mar 2003)
11. CZ.NIC: Czechs refused diacritics in domain names again (Feb 2017), https://www.nic.cz/page/3499/czechs-refused-diacritics-in-domain-names-again/
12. Davis, M., Suignard, M.: Unicode IDNA compatibility processing. Technical Standard 46, The Unicode Consortium (May 2018), https://www.unicode.org/reports/tr46/
13. DENIC: DENIC putting extensive changes into force for .DE Whois lookup service by 25 May 2018 (May 2018), https://www.denic.de/en/whats-new/press-releases/article/denic-putting-extensive-changes-into-force-for-de-whois-lookup-service-as-of-25-may-2018/
14. Dhamija, R., Tygar, J.D., Hearst, M.: Why phishing works. In: SIGCHI Conference on Human Factors in Computing Systems. pp. 581–590. ACM (2006). https://doi.org/10.1145/1124772.1124861
15. Dinaburg, A.: Bitsquatting: DNS hijacking without exploitation. White Paper #2011-307, Raytheon Company (2011)
16. Edelman, B.: Large-scale registration of domains with typographical errors. Tech. rep., Berkman Center for Internet & Society - Harvard Law School (Sep 2003), http://cyber.law.harvard.edu/people/edelman/typo-domains
17. Eskandari, S., Leoutsarakos, A., Mursch, T., Clark, J.: A first look at browser-based cryptojacking. In: 3rd IEEE European Symposium on Security and Privacy Workshops – Security on Blockchains. pp. 58–66 (2018). https://doi.org/10.1109/EuroSPW.2018.00014
18. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: 2nd International Conference on Knowledge Discovery and Data Mining. pp. 226–231. AAAI Press (1996)
19. EURid, UNESCO: World report on internationalised domain names 2018 (Aug 2018), https://idnworldreport.eu/2018-2

20. Faltstrom, P., Hoffman, P., Costello, A.: Internationalizing domain names in applications (IDNA). RFC 3490 (Mar 2003)
21. Gabrilovich, E., Gontmakher, A.: The homograph attack. Communications of the ACM **45**(2), 128 (Feb 2002). https://doi.org/10.1145/503124.503156
22. GoDaddy: The GoDaddy API, https://developer.godaddy.com/
23. Google: Safe browsing, https://safebrowsing.google.com/
24. Hannay, P., Baatard, G.: The 2011 IDN homograph attack mitigation survey. In: International Conference on Security and Management. pp. 653–657 (2012)
25. Hannay, P., Bolan, C.: An assessment of internationalised domain name homograph attack mitigation implementations. In: 7th Australian Information Security Management Conference (2009). https://doi.org/10.4225/75/57b405aa30dee
26. Hannay, P., Bolan, C.: The 2010 IDN homograph attack mitigation survey. In: International Conference on Security and Management. pp. 611–614 (2010)
27. Harrenstien, K., Stahl, M., Feinler, E.: DoD Internet host table specification. RFC 952 (Oct 1985)
28. Holgers, T., Watson, D.E., Gribble, S.D.: Cutting through the confusion: A measurement study of homograph attacks. In: USENIX Annual Technical Conference. pp. 261–266. USENIX Association (2006)
29. IDN Guidelines Working Group: Guidelines for the implementation of internationalized domain names, version 4.0 (May 2018), https://www.icann.org/en/system/files/files/idn-guidelines-10may18-en.pdf
30. Internet Assigned Numbers Authority: Repository of IDN practices, https://www.iana.org/domains/idn-tables
31. Internet Corporation for Assigned Names and Numbers: Label Generation Rules Tool, https://www.icann.org/resources/pages/lgr-toolset-2015-06-21-en
32. Internet Corporation for Assigned Names and Numbers: Data protection/privacy issues (Jul 2017), https://www.icann.org/dataprotectionprivacy
33. Kharraz, A., Robertson, W., Kirda, E.: Surveylance: Automatically detecting online survey scams. In: 39th IEEE Symposium on Security and Privacy. pp. 70–86 (2018). https://doi.org/10.1109/SP.2018.00044
34. Kintis, P., Miramirkhani, N., Lever, C., Chen, Y., Romero-Gómez, R., Pitropakis, N., Nikiforakis, N., Antonakakis, M.: Hiding in plain sight: A longitudinal study of combosquatting abuse. In: 24th ACM SIGSAC Conference on Computer and Communications Security. pp. 569–586. ACM (2017). https://doi.org/10.1145/3133956.3134002
35. Klensin, J.: Internationalized domain names for applications (IDNA): Definitions and document framework. RFC 5890 (Aug 2010)
36. Korczyński, M., Wullink, M., Tajalizadehkhoob, S., Moura, G.C.M., Noroozian, A., Bagley, D., Hesselman, C.: Cybercrime after the sunrise: A statistical analysis of DNS abuse in new gTLDs. In: 13th Asia Conference on Computer and Communications Security. pp. 609–623. ACM (2018). https://doi.org/10.1145/3196494.3196548
37. Krawetz, N.: Looks like it (May 2011), https://www.hackerfactor.com/blog/index.php?/archives/432-Looks-Like-It.html
38. Larsen, C., van der Horst, T.: Bad guys using internationalized domain names (IDNs) (May 2014), https://www.symantec.com/connect/blogs/bad-guys-using-internationalized-domain-names-idns
39. Le Pochat, V., Van Goethem, T., Tajalizadehkhoob, S., Korczyński, M., Joosen, W.: Tranco: A research-oriented top sites ranking hardened against manipulation. In: 26th Annual Network and Distributed System Security Symposium (Feb 2019). https://doi.org/10.14722/ndss.2019.23386

40. Levine, J., Hoffman, P.: Variants in second-level names registered in top-level domains. RFC 6927 (May 2013)
41. Liu, B., Lu, C., Li, Z., Liu, Y., Duan, H., Hao, S., Zhang, Z.: A reexamination of internationalized domain names: The good, the bad and the ugly. In: 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks. pp. 654–665 (2018). https://doi.org/10.1109/DSN.2018.00072
42. Liu, S., Foster, I., Savage, S., Voelker, G.M., Saul, L.K.: Who is .com?: Learning to parse WHOIS records. In: Internet Measurement Conference. pp. 369–380. ACM (2015). https://doi.org/10.1145/2815675.2815693
43. Lv, P., Ya, J., Liu, T., Shi, J., Fang, B., Gu, Z.: You have more abbreviations than you know: A study of AbbrevSquatting abuse. In: 18th International Conference on Computational Science. pp. 221–233. Springer (2018)
44. Markham, G.: IDN display algorithm (Apr 2017), https://wiki.mozilla.org/IDN_Display_Algorithm
45. McElroy, T., Hannay, P., Baatard, G.: The 2017 homograph browser attack mitigation survey. In: 15th Australian Information Security Management Conference. pp. 88–96 (2017). https://doi.org/10.4225/75/5a84f5a495b4d
46. Mockapetris, P.: Domain names - concepts and facilities. RFC 1034 (Nov 1987)
47. Moore, T., Edelman, B.: Measuring the perpetrators and funders of typosquatting. In: 14th International Conference on Financial Cryptography and Data Security. pp. 175–191. Springer (2010)
48. Nikiforakis, N., Balduzzi, M., Desmet, L., Piessens, F., Joosen, W.: Soundsquatting: Uncovering the use of homophones in domain squatting. In: 17th International Conference on Information Security. pp. 291–308. Springer (2014)
49. Nikiforakis, N., Maggi, F., Stringhini, G., Rafique, M.Z., Joosen, W., Kruegel, C., Piessens, F., Vigna, G., Zanero, S.: Stranger danger: Exploring the ecosystem of ad-based URL shortening services. In: 23rd International Conference on World Wide Web. pp. 51–62. ACM (2014). https://doi.org/10.1145/2566486.2567983
50. Nikiforakis, N., Van Acker, S., Meert, W., Desmet, L., Piessens, F., Joosen, W.: Bitsquatting: Exploiting bit-flips for fun, or profit? In: 22nd International Conference on World Wide Web. pp. 989–998. ACM (2013). https://doi.org/10.1145/2488388.2488474
51. Nominet: .wales and .cymru domains - IDN policy (Aug 2015), https://nominet-prod.s3.amazonaws.com/wp-content/uploads/2015/08/CymruWalesIDNPolicy_0.pdf
52. Núcleo de Informação e Coordenação do Ponto BR: Regras do domínio, https://registro.br/dominio/regras.html
53. OpenDNS: PhishTank, https://www.phishtank.com
54. Rüth, J., Zimmermann, T., Wolsing, K., Hohlfeld, O.: Digging into browser-based crypto mining. In: Internet Measurement Conference. pp. 70–76. ACM (2018). https://doi.org/10.1145/3278532.3278539
55. Scheitle, Q., Hohlfeld, O., Gamba, J., Jelten, J., Zimmermann, T., Strowes, S.D., Vallina-Rodriguez, N.: A long way to the top: Significance, structure, and stability of Internet top lists. In: Internet Measurement Conference. pp. 478–493. ACM (2018). https://doi.org/10.1145/3278532.3278574
56. Schiffman, M.: Global internationalized domain name homograph report, Q2/2018. Tech. rep., Farsight Security (Jun 2018)
57. Shin, J.: Establish a process to update "top domain" skeleton list for confusability check (May 2017), https://bugs.chromium.org/p/chromium/issues/detail?id=722022

58. Shin, J.: Mitigate spoofing attempt using Latin letters (Apr 2017), https://codereview.chromium.org/2784933002
59. Sommers, J.: On the characteristics of language tags on the web. In: 19th International Conference on Passive and Active Measurement. pp. 18–30. Springer (2018). https://doi.org/10.1007/978-3-319-76481-8_2
60. Spamhaus Project: The domain block list, https://www.spamhaus.org/dbl/
61. Spaulding, J., Upadhyaya, S., Mohaisen, A.: The landscape of domain name typosquatting: Techniques and countermeasures. In: 11th International Conference on Availability, Reliability and Security. pp. 284–289 (2016). https://doi.org/10.1109/ARES.2016.84
62. SURBL: SURBL URI reputation data, http://www.surbl.org/
63. Szurdi, J., Kocso, B., Cseh, G., Spring, J., Felegyhazi, M., Kanich, C.: The long "taile" of typosquatting domain names. In: 23rd USENIX Security Symposium. pp. 191–206. USENIX Association (2014)
64. The Unicode Consortium: Unicode transliteration guidelines, http://cldr.unicode.org/index/cldr-spec/transliteration-guidelines
65. The Unicode Consortium: The Unicode Standard, Version 11.0.0 (2018), http://www.unicode.org/versions/Unicode11.0.0/
66. Tian, K., Jan, S.T.K., Hu, H., Yao, D., Wang, G.: Needle in a haystack: Tracking down elite phishing domains in the wild. In: Internet Measurement Conference. pp. 429–442. ACM (2018). https://doi.org/10.1145/3278532.3278569
67. Vissers, T., Barron, T., Van Goethem, T., Joosen, W., Nikiforakis, N.: The wolf of name street: Hijacking domains through their nameservers. In: 24th ACM SIGSAC Conference on Computer and Communications Security. pp. 957–970. ACM (2017). https://doi.org/10.1145/3133956.3133988
68. Vissers, T., Joosen, W., Nikiforakis, N.: Parking sensors: Analyzing and detecting parked domains. In: 22nd Annual Network and Distributed System Security Symposium. Internet Society (2015)
69. Wang, Y.M., Beck, D., Wang, J., Verbowski, C., Daniels, B.: Strider typo-patrol: Discovery and analysis of systematic typo-squatting. In: 2nd Workshop on Steps to Reducing Unwanted Traffic on the Internet. pp. 31–36. USENIX Association (2006)
70. Wood, P., Johnston, N.: Spammers taking advantage of IDN with URL shortening services (Feb 2011), https://www.symantec.com/connect/blogs/spammers-taking-advantage-idn-url-shortening-services
71. Zheng, X.: Phishing with Unicode domains (Apr 2017), https://www.xudongz.com/blog/2017/idn-phishing/