20 YEARS

OXFORD INFORMATION LABS

Разнообразие через локализацию:

Как страновые домены способствуют языковому разнообразию в интернете

Эмили Тейлор

Содержание

1. Обзор

Мировой интернет говорит на английском языке. В мире существует более 6000 языков[1], но английский стабильно остается самым популярным: он используется на 54% всех веб-страниц. Самые распространенные мировые языки, такие как арабский или хинди, на которых говорит более миллиарда человек, едва представлены в интернете[2].

Исследование 16,4 млн. доменных имен, управляемых членами CENTR, показало, что 76% из них адресуют на контент, изданный на национальном языке.

В честь своего 20-летия Ассоциация европейских регистратур страновых доменов верхнего уровня CENTR заказала Oxford Information Labs исследование с целью проверить гипотезу о том, что национальные регистратуры поддерживают ресурсы на национальных языках. Благодаря сотрудничеству с членами ассоциации Oxford Information Labs получила беспрецедентный доступ к файлам зоны и возможность провести языковой анализ 16,4 миллиона географических доменных имен.

Полные и ассоциированные члены CENTR совместно управляют 80% зарегистрированных имен в страновых доменах всего мира[3]. Многие из них были созданы в 1990-х, перед началом коммерциализации доменного рынка. Создание большинства из них было обусловлено духом и принципами эпохи раннего интернета, так как это определено RFC 1591: «Регистратуры страновых доменов верхнего уровня – это доверенные структуры, которым делегировано право управлять доменом, и работа которых направлена на служение интернет-сообществу[4]». Обязательство обслуживать интересы местного интернет-сообщества напрямую определяет их работу по поддержке национальных языков.

В соответствии с RFC 1591, регистратуры страновых доменов верхнего уровня – это доверенные структуры, которым делегировано право управление доменом, и работа которых направлена на служение интернет-сообществу.

Наше исследование показало, что в среднем 76% веб-контента, адресуемого страновыми доменами, опубликовано на национальном языке той страны или территории, которой этот домен принадлежит. На английском языке опубликовано 19% контента, еще 4% приходится на другие языки.

Для тех доменов верхнего уровня в нашем исследовании, где возможна регистрация интернационализированных доменных имен, или IDN (то есть доменных имен с диактрическими знаками или с использованием букв нелатинских алфавитов), доля контента на национальном языке выше (84%), а на английском языке – меньше (9%). Это согласуется с выводом о том, что IDN-домены способствуют языковому разнообразию в интернете[5].

1 https://en.unesco.org/news/unesco-launches-website-international-year-indigenous-languages-iyil2019

2 See W3Techs 'Usage of content languages for websites, 2019 https://w3techs.com/technologies/overview/content_language/all

3 CENTRStats Global TLD Report, April 2019 https://stats.centr.org/stats/global

4 https://www.rfc-editor.org/rfc/rfc1591.txt

5 https://idnworldreport.eu/

2. Методика

Большинство регистратур страновых доменов верхнего уровня не публикует в открытом доступе свой файл зоны. Специалисты, которые проводят исследования языкового разнообразия в интернете, обычно не имеют доступа к к этим данным.

Команда исследователей Oxford Information Labs в течение многих лет разрабатывала собственную методику языкового анализа веб-контента, которая используется для подготовки ежегодного Всемирного отчета по интернационализированным доменным именам EURid UNESCO. Отработка алгоритма на публично доступных общих доменах верхнего уровня (gTLD) помогла решить проблему анализа большого массива данных, собранных в разных форматах. Для этого исследования команда разработчиков усовершенствовала методику для эффективного определения одностраничных сайтов и доменов на паркинге.

В рамках 20-го юбилея CENTR ассоциация разослала по своим членам призыв к сотрудничеству. Ответ был положительным. Это исследование представляет результаты языкового анализа 10 страновых доменов верхнего уровня, в которых зарегистрировано 16,4 миллиона имен. Этот составляет 16% от всех доменных имен, управляемых полными и ассоциированными членами CENTR.

Данные были предоставлены в период с декабря 2018 года по май 2019.

Автоматический языковой анализ включал в себя следующие операции:

- Определение делегированных доменов.
- Автоматический языковой анализ активных доменов, вошедших в выборку.
- Определение страниц с низкоприоритетным контентом, в том числе одностраничных сайтов или страниц, размещенных на известных парковочных ресурсах.
- Сравнение результатов языкового анализа до и после исключения ресурсов.
- Определение интернационализированных доменных имен (IDN) и сравнение результатов, полученных для них, с результатами для полной выборки.

Более подробно методика исследования описана в приложении.

3. Какие домены верхнего уровня попали в исследование?

TLD	Страна или территория	Основной язык [6]	Количество доменов в зоне	Примечание
.cat	Каталония	Каталонский	0.1 млн.	Не страновой, а общий TLD, созданный для каталонского сообщества.
.ch	Швейцария	Немецкий, французский, итальянский	2.1 млн.	
.dk	Дания	Датский	1.3 млн.	
.nl	Нидерланды	Голландский	4.5 млн.	SIDN, регистратура странового домена .nl, также поделилась с командой исследователей результатами собственного детального исследования.
.nu	Ниуэ	Шведский	0.5 млн.	Поддерживается регистратурой национального домена Швеции.
.pt	Португалия	Португальский	0.3 млн.	Всего в домене .pt зарегистрировано 1.1 млн. имен [7]. DNS.pt предоставила часть файла зоны.
.ru / .рф	Российская Федерация	Русский	5.8 млн.	Вместе в доменах верхнего уровня .ru и .рф.
.se	Швеция	Шведский	1.4 млн.	
.sk	Словакия	Словацкий	0.4 млн.	
Общее количество исследованных имен			16.4 млн.	

Таблица 1. Исследованные домены верхнего уровня

6 Источник: Ethnologue

7 Информация изменена 22 мая 2019, источник https://www.dns.pt/pt/ estatisticas/

4. Результаты

4.1.1 Активные домены

Для того чтобы домен мог адресовать на какой-то контент, он должен быть активным. В среднем доля активных доменных имен (у которых указаны ns-записи или на которых созданы адреса электронной почты) в каждом домене верхнего уровня в нашем исследовании составляет 80%. Конкретные показатели варьируются в большом диапазоне: самый большой процент делегированных доменов оказался в зоне .sk (Словакия) – 91%, а самый маленький – в зоне .nu (Ниуэ) – 44%.

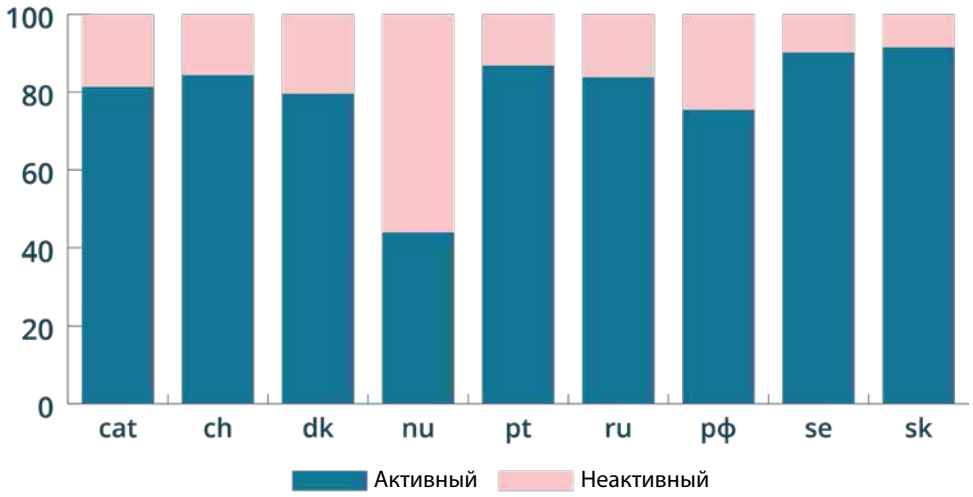


Рисунок 1. Доля активных доменов

Для IDN-доменов этот показатель составил в среднем 72%, самый низкий процент – у домена .cat (36%), самый высокий – у домена .pt (93%).

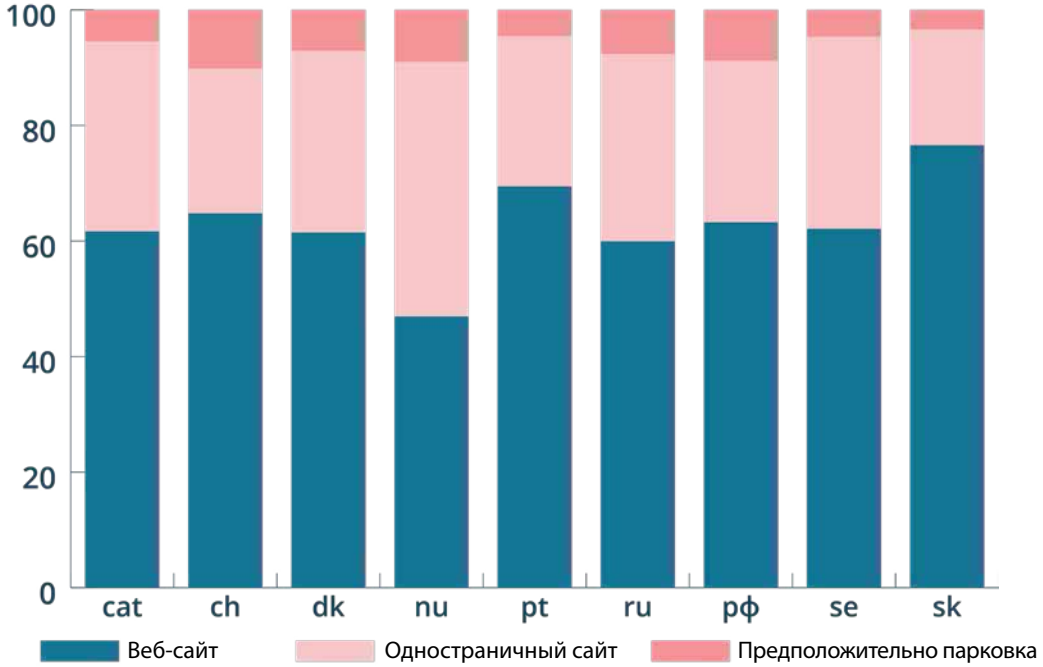
Для некоторых доменов результат может быть не репрезентативен, так как отдельные регистратуры предоставили только часть файла зоны, содержащую делегированные домены, или исключив из него непродленные имена, ожидающие удаления из реестра.

Задача этого исследования не оценить степень использования имен в каждом из доменов верхнего уровня, а исключить неделегированные домены из дальнейшего рассмотрения.

4.1.2 Использование доменных имен: определение низкоприоритетного контента

Регистрация доменного имени – это, как правило, только первый шаг на пути открытия нового бизнеса или интернет-проекта. Иногда требуются месяцы и годы для того, чтобы создать на нем полноценный сайт. В это время пользователь, используя сервисы регистратора или хостинг-провайдера, может разместить на домене одностраничный сайт или разместить домен на парковочном сервисе. И одностраничные сайты, и парковки обычно создаются типовыми, с одинаковым или очень похожим текстом на них. Парковочные страницы также часто содержат рекламу, чтобы монетизировать трафик.

Исследователи определили домены, на которых расположена информация такого рода, и исключили их из исследования. Подробнее о том, как определялись домены на паркинге, можно прочитать в приложении с описанием методики.



В среднем доля низкоприоритетного котента (одностраничные сайты и парковка) составила 37% для доменов в рассматриваемой выборке.

Рисунок 2. Использование имен в доменах верхнего уровня

В среднем количество низкоприоритетного контента составило 37%. Самый высокий процент оказался в домене .nu (53%), а самый низкий – в домене .sk (23%).

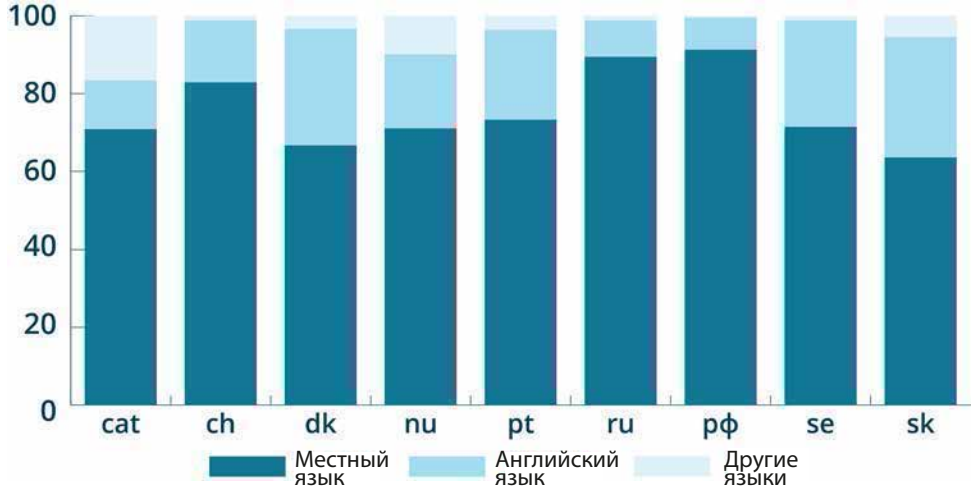
IDN-домены в рассматриваемой выборке (и первого, и второго уровней) содержат в среднем большее количество низкоприоритетного контента. Самый высокий уровень у домена .nu (58%), самый низкий – у домена .pt (2%). Домен .sk не поддерживает регистрацию имен на национальном языке.

4.1.3 Языковой анализ

После исключения доменов с низкоприоритетным контентом исследователи провели языковой анализ веб-контента, размещенного на оставшихся доменных именах из рассматриваемой выборки (в соответствии с описанной ниже методикой).

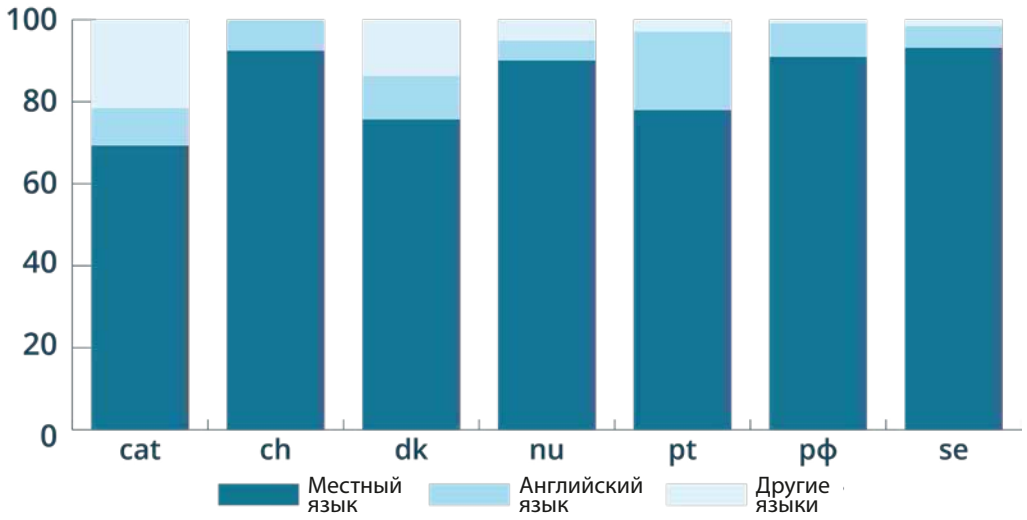
4.1.3.1 Использование национального языка

Во всех рассматриваемых доменных зонах доля основного языка соответствующей страны или территории составила, как минимум, 64% для всего проанализированного контента. Среднее значение составило 76%.



В среднем доля национального языка составила 76% для ресурсов в рассматриваемых доменах.

Рисунок 3. Результаты языкового анализа по страновым доменам (без учета низкоприоритетного контента)



Для IDN-доменов средняя доля контента на национальном языке составила 84%.

Рисунок 4. Языковой анализ для интернационализированных доменов верхнего уровня (без учета низкоприоритетного контента)

IDN-домены (и первого, и второго уровня) показали среднюю долю веб-контента на национальном языке 84%. При этом для некоторых доменов, таких как .ch, .nu, .se и .рф, этот показатель составил более 90%. Самая низкая доля контента на национальном языке в домене .cat – 69%.

4.1.3.2 Национальный домен Нидерландов

SIDN, регистратура странового домена Нидерландов .nl, предоставила исследователям результаты собственного языкового исследования зоны.

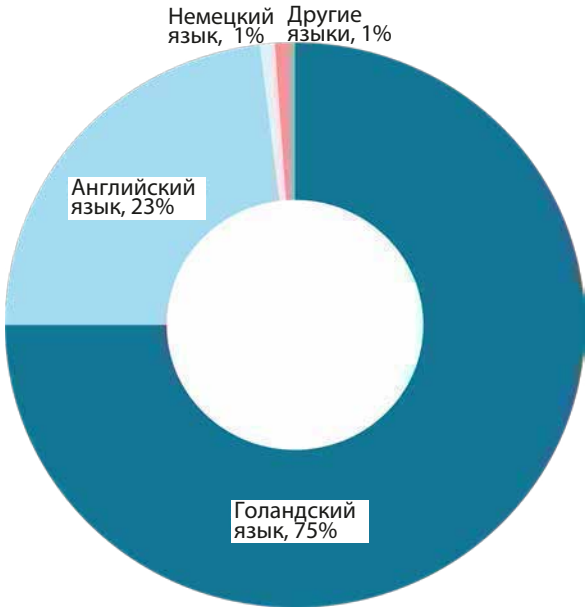


Рисунок 5. Результаты языкового исследования домена .nl (по информации SIDN)

Результаты исследования зоны .nl не противоречат общим результатам исследования: 75% контента, связанного с доменными именами в национальной доменной зоне Нидерландов, издано на голландском языке. Однако не известно, исключал ли SIDN из рассмотрения одностраничные сайты и домены, предположительно находящиеся на парковке.

4.1.3.3 Доля английского языка и влияние на нее ресурсов с низкоприоритетным контентом

Сравнение результатов языкового анализа до и после исключения из выборки имен, на которых расположен низкоприоритетный контент, показало, что количество ресурсов с контентом на английском языке существенно уменьшилось во всех страновых доменах.

	Вся зона		Только IDN	
TLD	До	После	До	После
.cat	50%	12%	58%	9%
.ch	37%	16%	30%	7%
.dk	48%	30%	41%	11%
.nu	50%	19%	42%	5%
.pt	44%	23%	19%	19%
.ru	26%	9%		
.рф (IDN)	16%	8%	16%	8%
.se	43%	27%	39%	8%
.sk	41%	31%		
СРЕДНИЙ ПОКАЗАТЕЛЬ	39%	19%	35%	9%

Таблица 2. Контент на английском языке до и после удаления из выборки ресурсов с низкоприоритетным контентом

В среднем доля контента на английском языке среди исследуемых доменов до удаления ресурсов с низкоприоритетным контентом составила 39%, при этом самый высокий показатель был в доменах .cat и .nu (50%), а самый низкий – в домене .ru (26%). После исключения из выборки ресурсов с такой информацией средняя доля контента на английском уменьшилась до 19%, при этом самое большое изменение наблюдалось в зоне .cat (с 50% до 12%), а самое маленькое – в зоне .sk (с 41% до 31%).

Английский язык с высокой вероятностью используется для ресурсов с низкоприоритетным контентом, таких как одностраничные сайты или парковки.

Среди IDN-доменов (и первого, и второго уровня) средняя доля контента на английском языке до удаления из выборки ресурсов с низкоприоритетным контентом составила 35%. Самая большая доля была в зоне .cat (58%), а самая маленькая – в зоне .pt. После исключения ресурсов с таким контентом из выборки средняя доля контента на английском языке упала до 9%. При этом самое большое изменение наблюдалось в зоне .cat (с 58% до 9%), а самое маленькое – в зоне .pt, где этот показатель остался неизменным.

4.1.3.4 Представленность малых языков

2019 год объявлен ЮНЕСКО Годом языков коренных народов[8], из которых 2860 находятся под угрозой исчезновения. Несколько языков, входящих в этот список, есть и в Европе: среди них корсиканский, галисийский, ирландский, валлийский и баскский. Современное интернет-окружение способствует использованию английского и других самых популярных языков. В то же время другие языки, находящиеся на грани исчезновения, представлены в интернете даже меньше, чем в реальной жизни.

Исследование показало незначительное присутствие малых языков в рассматриваемых доменных зонах.

Мы проверили, присутствует ли контент на малых (в особенности на исчезающих) языках в доменах верхнего уровня из нашей выборки. В частности, нам было интересно, присутствует ли саамский язык (на котором сейчас говорят около 30 000 человек в скандинавских странах[9]) в доменах .se или .nu, но это оказалось невозможным, так как этот язык пока не поддерживается сервисом Google translate. К сожалению, выборочная проверка других исчезающих языков показала ошибки инструментов автоматического перевода.

8 UNESCO international year of indigenous languages 2019 https://en.iyil2019.org/

9 UNESCO Atlas of the World's Languages in Danger, http://www.unesco.org/languages-atlas/index.php

5. Выводы и заключения

Исследование показывает, что домены верхнего уровня стран и регионов способствуют развитию контента на национальных языках. Доля контента на английском языке в этих доменах оказалась меньше, чем в среднем по миру.

Домены верхнего уровня стран и регионов способствуют развитию в интернете контента на национальных языках.

Языковые паттерны разных доменных зон не случайны, а в целом соответствуют основному языку страны или территории домена. Так, в зоне .sk 64% (91 000+) ресурсов содержат контент на словацком языке, в то время как среди всех мировых сайтов он занимает только 0,4%[10]. При этом доля «чужих» языков в каждой из таких зон сравнительно невелика и не превышает 5%. Это свидетельствует о том, что национальные домены верхнего уровня, в первую очередь, представляют в интернете свою страну и ее язык.

Таблица 3. Результаты исследования каждого из рассматриваемых страновых доменов.

Домен	Национальный язык страны или территории	Доля делегированных доменов	Парковка+	Основной язык	2-й популярный язык	3-й популярный язык*	Доля других языков
.cat	Каталонский	81%	38%	Каталонский	Испанский	Английский	4%
.cat IDNs		36%	53%				0%
.ch	Немецкий, французский, итальянский, романшский	84%	35%	Немецкий	Английский	Французский	1%
.ch IDNs		77%	43%				1%
.dk	Датский	80%	39%	Датский	Английский	Шведский	3%
.dk IDNs		52%					3%
.nl	Голландский	Нет данных		Голландский	Английский	Немецкий	1%
.nu	Шведский, датский, голландский	44%	53%	Шведский	Английский	Голландский	5%
.nu IDNs		86%	58%			Датский	3%
.pt	Португальский	87%	31%	Португальский	Английский	Испанский	2%
.pt IDNs		93%	2%				2%
.ru	Русский	84%	40%	Русский	Английский	Болгарский	<2%
.рф		75%	37%				<1%
.se	Шведский	90%	38%	Шведский	Английский	Немецкий	<2%
.se IDNs		86%	55%	Шведский	Английский	Румынский	<1%
.sk	Словацкий	91%	23%	Словацкий	Английский	Чешский	3%

+ Парковка включает в себя одностраничные сайты и домены, предположительно находящиеся на парковке (см. главу 4)

*3-й популярный язык включен в «остальные» языки в графиках в главе 4.

10 W3Techs, 2019 https://w3techs.com/technologies/overview/content_language/all

Три самых популярных языка, представленных в каждом из страновых доменов (исключая английский), совпадают с наиболее популярными языками в стране, которой этот домен принадлежит. Второй и третий по популярности языки также обусловлены географическим расположением страны и тем, что они принадлежат к одной языковой семье. Так, испанский язык представлен в зонах .cat и .pt, чешский – в зоне .sk, шведский – в зоне .dk. Единственное исключение – это присутствие румынского языка в IDN-доменах второго уровня зоны .se, но доля его невелика (меньше 1%).

Английский язык занимает заметные позиции в каждом из исследуемых доменов, но во всех случаях его доля значительно меньше среднего по миру показателя в 54%[11].

Использование английского языка оказалось существенно ниже, чем в среднем по миру.

Языковой анализ IDN-доменов показал меньшую долю английского языка и большую представленность национальных языков, чем в общей выборке.

После удаления ресурсов с низкоприоритетным контентом (доменов, предположительно находящихся на парковке или ведущих на одностраничные сайты), доля контента на английском языке уменьшилась во всех исследуемых доменах. Это позволяет сделать вывод, что подобная информация чаще всего публикуется на английском языке.

В то время как страновые домены показывают четкую корреляцию с титульными языками соответствующей страны или территории, присутствие местных (или малых) языков остается очень слабым. Несмотря на глобальную природу интернета, использование языков онлайн не соответствует картине, наблюдаемой в реальном мире. В этой связи роль регистратур страновых доменов в поддержке языкового разнообразия представляется особенно важной.

Усилия национальных регистратур по поддержке языкового разнообразия в интернете редко получают достойную оценку, так как в открытых источниках крайне мало данных об использовании языков, и большая часть регистратур не делает свои файлы зоны открытыми для исследований. Следовательно, проводится очень мало исследований, какие языки используются в национальных доменах. Специалисты Oxford Information Labs благодаря сотрудничеству с членами CENTR смогли получить доступ к данным о более чем 16,4 миллионах доменных имен и определить, насколько домены стран и территорий способствуют использованию национальных языков в интернете.

11 ibid

Приложение: методика исследования

Сбор данных

Мы предложили полным и ассоциированным членам ассоциации CENTR поделиться своими файлами зоны или результатами собственных языковых исследований.

Команда исследователей получила данные для анализа по девяти доменам .ch, .dk, .se и .nu, .pt, .ru и .рф, .sk. Помимо этого мы получили доступ к реестру домена .cat (общий домен верхнего уровня Каталонии) через централизованный информационный сервис (CZDS File) корпорации ICANN. В общей сложности исследуемые зоны содержат 12 миллионов записей. Также регистратура Нидерландов SIDN привела результаты собственного языкового исследования 4,5 миллионов доменных имен в зоне .nl. Таким образом, исследование охватывает в общей сложности 16,4 миллиона доменных имен, что составляет 16% от всех доменов, находящихся под управлением полных и ассоциированных членов ассоциации CENTR.

Регистратуры предоставляли сведения о своих доменах по различным правилам. Файлы зон некоторых из них, таких как .se или .nu, находятся в свободном доступе. Другие, например, .pt, предоставили данные только о доменах, которые давали какой-то ответ на DNS-запросы. В результате общее количество доменных имен, по которым проводилось исследование, может не совпадать с количеством имен, которые находятся под управлением регистратур из приведенной выборки. Методика языкового исследования SIDN также может отличаться от методики, по которой работали специалисты Oxford Information Labs.

Данные для исследований были предоставлены в период с декабря 2018 по май 2019 года.

Анализ данных

Определение активных доменов

Исследователи выполнили следующие операции с каждым доменным именем из списка:

- Проверили, существуют ли для доменного имени записи NS или MX. Если для домена существовали записи A или AAAA, NS и MX, то проверялось наличие записей для домена третьего уровня www на этом доменном имени.
- Удалили из выборки домены, у которых не было ни одной вышеперечисленной записи, и присвоили им статус «неактивный».

Эти шаги позволили нам выявить активные домены в каждой из доменных зон. Результаты могут быть искажены, если регистратура передала список только активных доменов. Это могло бы вызвать искажение в результатах исследования для таких регистратур в пользу активных доменов.

Для оставшихся доменов проводились следующие операции:

- Проверили, существует ли на этом домене какой-либо веб-контент (статус домена и наличие главной страницы).
- Определили, настроена ли на домене переадресация, и собрали такие домены в отдельный список.
- Проверили, можно ли отнести домен к ресурсам с низкоприоритеным контентом по следующим критериям: незначительное количество внутренних ссылок (определение одностраничных сайтов), наличие повторяющегося контента, наличие ссылок, которые ведут на одну страницу из многих источников, количество слов в тексте более 50, наличие текста на латыни (для тестовых страниц часто используется латинский шаблон Lorem ipsum).

Языковой анализ для всех делегированных доменов

Автоматический языковой анализ включал следующие операции:

- Анализ сохраненного текста и выявление десяти наиболее часто встречающихся слов, игнорируя неинформативные.
- Перевод ключевых слов при помощи автоматизированных инструментов.
- Исключение возможных ошибок автоматизированного перевода, повторный прогон через переводческие инструменты целых фраз из текста.

Повторный языковой анализ списка доменов, из которых исключены имена, ведущие на одностраничные сайты или предполагаемые парковки.

После исключения из выборки доменов, ведущих на одностраничные сайты или сайты, предположительно находящиеся на парковке, языковой анализ был проведен еще раз.

Определялись «местные» языки для каждого из доменов верхнего уровня (см. таблицу 1).

Анализ IDN- доменов

Определение IDN-доменов в выборке по начальным символам имени xn--, затем повторение пунктов с 2.2.1 до 1.2.3

Примечание

Команда исследователей тщательно работала над результатами анализа, но ошибки случаются. Средства автоматического перевода, которые применялись для этой работы (Google translate), обычно работают хорошо, но иногда допускают неточности. В случаях, когда одно и то же слово встречается в разных языках (например, в шведском и датском или словацком и чешском) автоматический переводчик мог спутать один язык с другим. Инструмент также не применим при анализе редких языков, таких как галисийский, корсиканский, ирландский или валлийский, поскольку не распознает языки, находящиеся под угрозой исчезновения.

Мы не можем быть полностью уверены, что домены в зоне .pt, контент на которых был определен как испаноязычный, в действительности является таковым. Ручная проверка выборочных страниц показала, что некоторые страницы на самом деле были на португальском. Также и некоторые тексты на доменах .sk, которые были определены как чешские, при ручной проверке оказались словацкими. Исследователи провели повторную проверку целых фраз из этой выборки при помощи автоматических переводчиков, но ошибки, вероятно, все еще остались.

Ассоциация европейских регистратур национальных доменов верхнего уровня

CENTR – это Ассоциация европейских регистратур национальных доменов верхнего уровня, таких как .de для Германии или .si для Словении. В настоящее время в CENTR входят 54 полных и 9 ассоциированных членов, вместе они управляют более 80% всех мировых доменных имен. Задачи Ассоциации – разрабатывать и продвигать среди регистратур национальных доменов лучшие отраслевые практики. Статус полного члена CENTR доступен организациям, коммерческим компаниям или частным лицам, которые выполняют функции регистратуры национального домена верхнего уровня.

Этот документ является частью серии работ, посвященных отраслевым исследованиям, анализу исторических данных и прогнозам развития технологий, таких как цифровая идентификация, опубликованных в течение 2019 года и приуроченных к 20-й годовщине Ассоциации. Точка зрения авторов этих публикаций может не совпадать с точкой зрения Ассоциации или ее членов.

CENTR выражает признательность и благодарит за помощь в проведении 20-го юбилея Ассоциации следующие организации:

Платиновый спонсор

Золотые спонсоры

Серебряные спонсоры

CENTR vzw/asbl
Belliardstraat 20 (6th floor)
1040 Brussels, Belgium
Tel: +32 2 627 5550
Fax: +32 2 627 5559
secretariat@centr.org
www.centr.org

Подпишитесь на нас в Twitter, Facebook или LinkedIn чтобы быть в курсе последних новостей