2003 - Том 6 - Выпуск 3

# Некоторые тенденции в эволюции общедоступного веба

# 1998 – 2002

*Edward T. O'Neill*, Brian F. Lavoie, Rick Bennett

**OCLC Office of Research**

## Введение

Быстрота, с которой веб превратился из технической диковинки в феномен культуры, поистине удивительна. Менее чем за десять лет он проник во все сферы общества, от коммерции до образования и используется в самых разных ситуациях, от школьных исследований до поиска прецедентов. Как и в случае других трансформирующих технологий-предшественников, веб породил множество последствий. Недавняя ситуация с технологиями «точка-ком» ("dot-com bust") ясно показал, что природа и возможности веба все еще осознанные в недостаточной степени.

Несмотря на то, что веб активно развивается, уже накопилось достаточно материала для содержательного анализа тенденций его эволюции. Короткая история веба сопровождалась огромным количеством предсказаний о дальнейшем направлении его развития, а также о той роли, которую он обязательно будет играть в качестве коммуникационной цифровой среды. В свете той перманентной неопределенности, которая сопровождает становление веба, полезно проанализировать некоторые тенденции в его современном развитии – как для того, чтобы выделить основные черты настоящего, так и для того, чтобы высказать некоторые новые предположения о его будущем.

В настоящей статье рассматриваются три тенденции в развитии публичного веба: размер и рост; интернационализация; использования метаданных. Данные для такого анализа получены в ходе проекта исследовательского отдела OCLC «Характеристики веба» (OCLC Office of Research Web Characterization Project) [1], - инициативы, которая посвящена исследованию фундаментальных проблем веба и его контента в ходе серии т.н. «Web samples», проводимых ежегодно с 1998.

## II. Характеристики публичного веба

В 1997 исследовательский отдел OCLC инициировал проект, призванный дать ответы на фундаментальные вопросы о вебе: насколько он велик? что он содержит? как он эволюционирует? Целью проекта было развитие и имплементация методологии для охарактеризования размера, структуры и

контента веба и предоставление получаемых результатов библиотечному сообществу и всем желающим.

Стратегия, принятая для охарактеризации веба, состояла в отборе репрезентативных образцов веб-сайтов для использования в качестве исходных данных для выведения оценок, которые корректно использовать в отношении всего веба в целом [2]. С использованием специально сконфигурированного генератора случайных чисел из всего пространства адресов 32-битного протокола IPv4 были случайным образом выбраны 0.1% IP-адресов. По каждому из этих адресов была предпринята попытка HTTP-соединения с 80-м портом (стандартным для веб-сервисов). IP-адрес признавался идентифицирующим веб-сайт, если в ответ на такую попытку возвращался ответ «code 200» и соответствующий документ. Затем контент веб-сайта собирался с помощью программного обеспечения, разработанного в OCLC.

После формирования такой коллекции веб-сайтов, были выполнены тесты с целью выявления сайтов-дубликатов, использующих несколько IP-адресов. В результате получалась оценка общего числа уникальных веб-сайтов. Затем, этот набор анализировался в целью выделения публичных сайтов. Под таким сайтом понимался такой, который предлагает всем пользователям веба свободный неограниченный доступ к значительной части своего контента. Публичный сайты могут содержать закрытые зоны, но для того, чтобы его можно было отнести к публичным, он должен содержать некоторое количество нетривиального контента, доступного без ограничений.

В результате такой процедуры была получена репрезентативная подборка уникальных публичных веб-сайтов. Набор всех публичных сайтов мы называем «публичный веб» - это та часть всего веба, которая доступна для обозрения и доступа среднего пользователя. Именно публичный веб является предметом анализа и обсуждения в данной статье.

After a pilot survey in 1997, the project conducted five consecutive surveys (1998 - 2002) of the Web, based on the sampling methodology described above.

## III. Размер и рост

Согласно данным последнего обзора проекта «Характеристики веба» (июнь 2002) публичный веб содержит 3,080,000 веб-сайтов или 35% всего веба. Публичные сайты содержат около 1.4 миллиарда веб-страниц. Средний размер сайта составил 441 страниц.

Насколько это значительные цифры? Недавно Shapiro и Varian [3] оценили общее число статических HTML страниц веба эквивалентными 1.5 миллиону книг. Эту цифру они сравнили с количеством томов в библиотеке университета Калифорнии в Беркли (8 миллионов) и – принимая во внимание тот факт, что только часть содержащейся в вебе информации можно рассматривать как полезную, - заключили, что веб «не слишком впечатляет как информационный ресурс».

Однако оценки Shapiro и Varian выглядят неточными. Цифровые ресурсы веба

представляют собой не только гладкий текст, но и многое другое, включая комбинированные мультимедийные информационные объекты. Оценка размеров веба только по статическим текстам ведет к игнорированию большей части содержащейся в нем информации. Далее, многие веб-аналитики признают различие между «поверхностным вебом» ("surface Web") и «глубинным вебом» ("deep Web"). Хотя такая терминология страдает неточностью и многозначностью, поверхностный веб может интерпретироваться как та его часть, которая доступна через гиперссылки с использованием обычных технологий сбора информации (crawling technologies). Такой подход используется в большинстве поисковых машин при генерации их индексов. Глубинный веб состоит из информации, которая недоступна для поисковых роботов по гиперссылкам: в частности, динамически генерируемые страницы, создаваемые в ходе взаимодействия между сайтом и пользователем. Например, онлайновые базы данных, создающие страницы в соответствии с параметрами запроса, являются частью глубинного веба. Хотя надежные оценки его размеров недоступны, предполагается, что они велики и быстро растут [4].

В другом исследовании Varian и Lyman [5] в 2000 году оценили поверхностный веб в 25 - 50 терабайт информации (исходя из допущения, что средний размер веб-страницы составляет 10-20 кб. Однако, авторы не проводили различия между публичными и иными видами веб-сайтов. Комбинируя данные из этой работы с результатами из обзора 2000 года проекта «Характеристики веба» и исходя из предположения, что веб-сайты всех типов имеют одинаковое число страниц, мы оцениваем публичный веб в 2000 году в 41% от поверхностного веба (10-20 терабайт). Для сравнения, в 2002 году поверхностный веб оценивается в 14-28 терабайт (по данным обзора 2002 года, считая средний размер веб-страницы в 10-20 кб).

Varian и Lyman оценивали место, занимаемое книгой из 300 страниц текста, в 1 Мб дискового пространства. Это означает, что в июне 2002 года поверхностный веб был примерно эквивалентен 14-28 миллионам книг. Такую оценку можно рассматривать как подтверждение вывода о том, что веб представляет собой коллекцию информации значительных размеров. Для сравнения, в 2001 году общее число книг в библиотеках 112 членов Ассоциации исследовательских библиотек (Association of Research Libraries, ARL) составляло около 3,7 миллионов экземпляров [6]. В библиотеке Гарвардского университета хранится около 15 миллионов книг.

Однако, вывод о том, что веб эквивалентен коллекции крупной библиотеки, был бы преждевременным. Значительный объем веба занимает разметка HTML и XML. Кроме того, Shapiro и Varian указывают на то, что большая часть информации может быть недоступна (is not generally useful cannot be dismissed lightly).

Наиболее замечательным аспектом является та скорость, с которой веб превратился из незначительного ресурса в информационную систему, по крайней мере, сравнимую с традиционной исследовательской библиотекой. Широко цитируемая работа [7] оценивала весь веб в 1996 в 100,000 сайтов. Два года спустя первый обзор проекта «Характеристики веба» оценил размеры только публичного веба в 1,5 миллиона сайтов. В 2000 году публичный веб вырос до 2,9

миллионов, а спустя два года, в 2002 году - до 3 миллионов сайтов. За пять лет проведения обзоров (1998 - 2002) публичный веб удвоился.

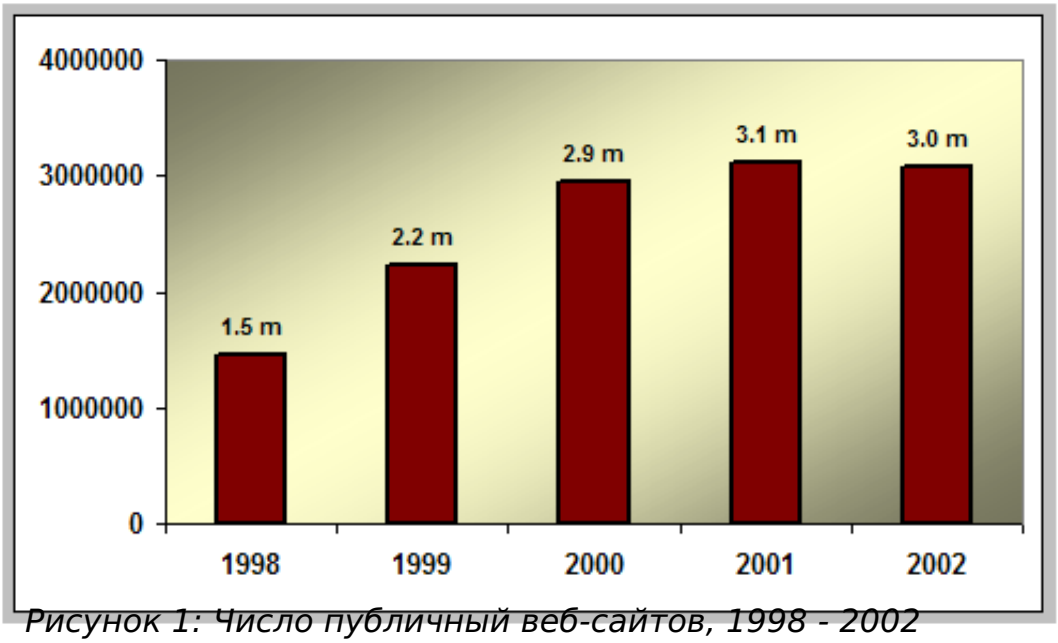


*Рисунок 1: Число публичный веб-сайтов, 1998 - 2002*

Однако этот впечатляющий рост маскирует важную тенденцию – темпы увеличения публичного веба существенно замедлились в конце указанного периода. Так, в 1998 и 1999 годах веб увеличивался на 50%, в 2000-2001 года рост сократился до 6%, а в 2001-2002 годах публичный веб даже несколько сократился в размерах. Т.е. большая часть роста веба, отмеченная в обзорах, пришлась на первые три года (1998 - 2000). В 1998 публичный веб составлял половину от величины 2002 года, однако в 2000 году его размер составлял 96% от этой же величины.

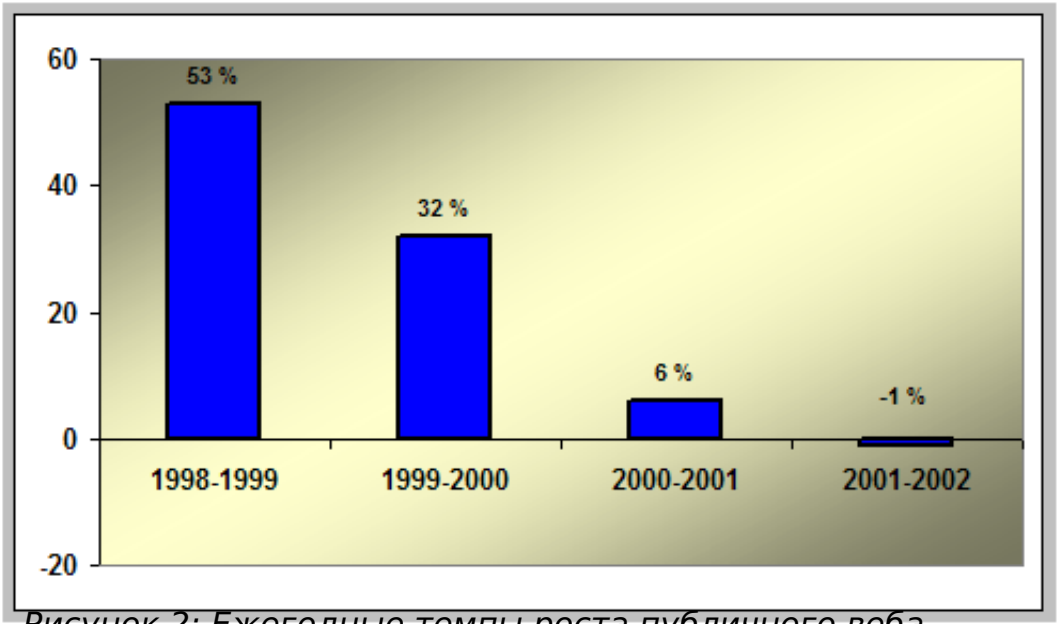


*Рисунок 2: Ежегодные темпы роста публичного веба*

Замедление темпов роста публичного роста еще более отчетливо видно из абсолютных цифр. С 1998 по 1999 год число сайтов выросло на 772,000; близкая

величина (713,000) добавилась за 1999-2000 годы. Однако, затем абсолютный рост резко снизился: за 2000-2001 года добавилось лишь 177,000 новых публичных веб-сайтов. А за 2001-2002 годы публичный веб сократился на 39,000 сайтов.

Таким образом, налицо стагнация, если не вовсе прекращение роста. Какими факторами это можно объяснить? Одна ключевая причина заключается в том, что веб больше не является новой технологией: те, кто хотел учредить свое присутствие в вебе – уже сделали это. В этом смысле лавинообразный переход в «онлайн», характерный для первых лет развития веба, сменился этапом поддержания и развития существующих сайтов. Действительно, данные обзора от июня 2002 года указывают, что при некотором уменьшении числа сайтов средний их размер становится больше. В 2001 году среднее число страниц на сайте равнялось 413; в 2002 году оно выросло до 441.

Наряду с уменьшением скорости появления новых сайтов, несколько увеличилась скорость исчезновения существующих. Анализ данных за 2001-2 годы показывает, что в 2002 году исчезло 17% из тех сайтов, что существовали в предыдущем году. Многие из тех, кто создавал веб-сайты в прошлом, пришли к выводу, что их поддержание более нецелесообразно. Ключевой фактор здесь – экономический. Многие фирмы просто ушли из онлайнового бизнеса, многие сократили объем операций в сети [8]. Другие аналитики отмечают снижение количества персональных веб-сайтов, принадлежащих отдельным людям. Среди причин называют истечение сроков бесплатных соглашений на веб-хостинг – люди не хотят платить за свои сайты [9].

Суммарно, наблюдаемый эффект замедления роста публичного веба объясняется двумя основными факторами: уменьшение числа новых сайтов и исчезновение существующих. Однако, это не означает, что публичный веб сокращается, по некоторым данным (например, число страниц или число терабайт) он фактически увеличивается. Но численные данные (а число сайтов приблизительно эквивалентно числу организаций, индивидуальных пользователей и коммерческих фирм, присутствующих в публичном вебе) указывают на то, что развитие вышло на плато.

## IV. Интернационализация

Как это следует из его названия, всемирная паутина представляет собой глобальный информационный ресурс в том смысле, что любой пользователь, независимо от страны или языка, может сделать свою информацию доступной в вебе. В таком случае, контент веба должен отражать международное сообщество в целом, включая источники из всех точек мира на большом числе языков.

В 1999 году была изучена принадлежность сайтов, попавших в репрезентативную выборку. Оказалось, что их владельцы (пользователи, организации или коммерческие фирмы) находились в 76 различных странах, что подтверждает глобальный характер сетевого сообщества. Однако, более тщательный анализ указывает на неоднозначность такого вывода. Реально, приблизительно половина публичных веб-сайтов принадлежало владельцам, находящимся в США. Ни одна другая страна не владела более 5% публичных сайтов и лишь только 8 стран

кроме США имели более 1%. Очевидно, что в 1999 году веб представлял собой США-центрированное информационное пространство.

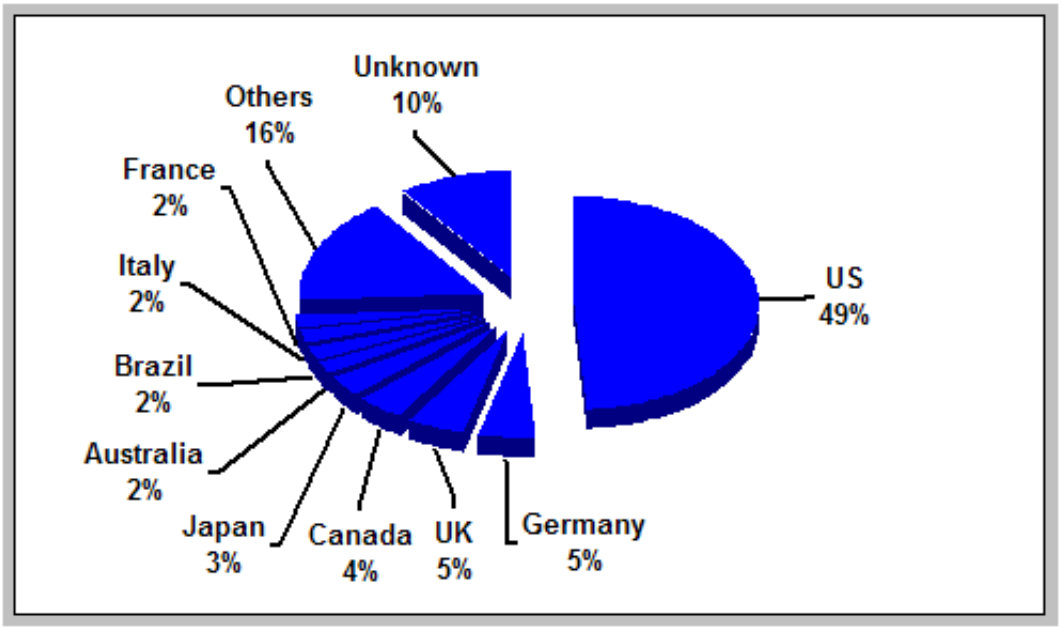


*Рисунок 3: Распределение публичных веб-сайтов по странам, 1999*

Мало что изменилось три года спустя. Доля американских сайтов в 2002 году даже несколько выросла (до 55%), хотя пропорции между остальными лидирующими странами остались примерно одинаковыми. В 2002, как и в 1999 году в выборку попали сайты из 76 стран – такой результат указывает, что веб не демонстрирует сколько-нибудь заметной тенденции к дальнейшей интернационализации.

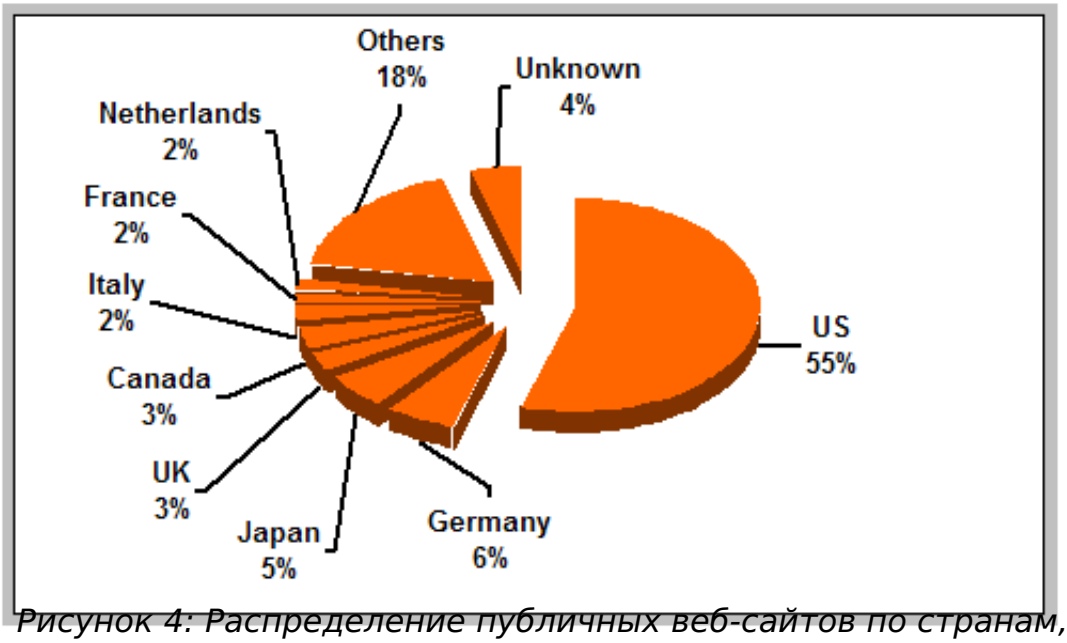


*Рисунок 4: Распределение публичных веб-сайтов по странам, 2002*

Такое заключение подкрепляется данными о языках текстового контента. Зная, что более половины публичных сайтов находится в США, легко подсказать, что английский язык будет превалировать в вебе. Однако в какой именно степени он доминирует? Как этот процесс эволюционирует?

Обзоры 1999 и 2000 годов содержат ответы на эти вопросы. В 1999 году были на сайтах, попавших в выборку, были идентифицированы 29 различных языков. На первый взгляд, это указывает на разнообразие веб-ресурсов с точки зрения языков их выражения. Однако, как и в случае с географией сайтов, эти данные переоценивают реальную степень интернационализации. Данные 1999 года указывают, что около 75% всех сайтов значительную часть своего контента выражали английскими языком. Следующим по частоте был немецкий язык, который встретился на 7% сайтов. Кроме английского, еще только 7 языков присутствовали на 2 и более процентах сайтов, включенных в выборку.

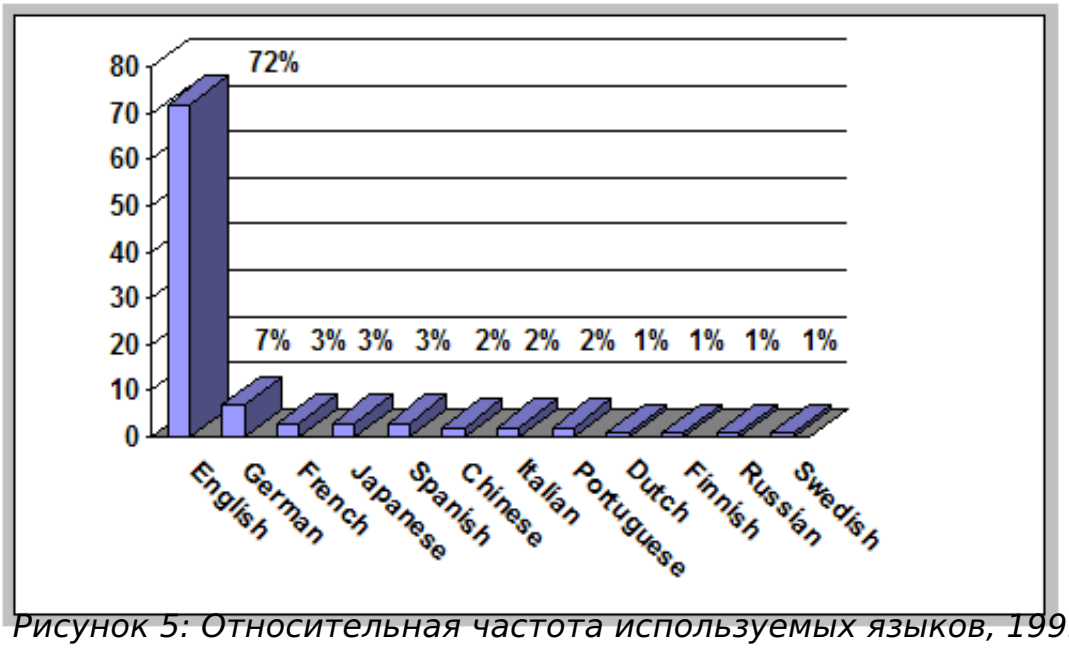


*Рисунок 5: Относительная частота используемых языков, 1999 (в процентах от общего числа публичных сайтов)*

Не все сайты представляли контент только на одном языке. Например, в 1999 году 7% публичных веб-сайтов содержали текстовую информацию на нескольких языках. Интересно, что во всех без исключения случаях одним из языков был английский.

Как и география публичных веб сайтов, распределение текстового контента по языкам за период с 1999 по 2002 годы изменилось незначительно. Доля публичных сайтов, предоставляющих значительную часть контента на английском по-прежнему около 75%. Ни один из других языков не превышает доли в 7%. Процент многоязычных сайтов несколько снизился: с 7 до 5%. Наиболее значительно выросло число сайтов на японском языке – с 3 до 6% за тот же период. Учитывая увеличение числа сайтов из Японии с 3 до 5%, можно сделать вывод, что японское присутствие в вебе существенно увеличилось за последние несколько лет.

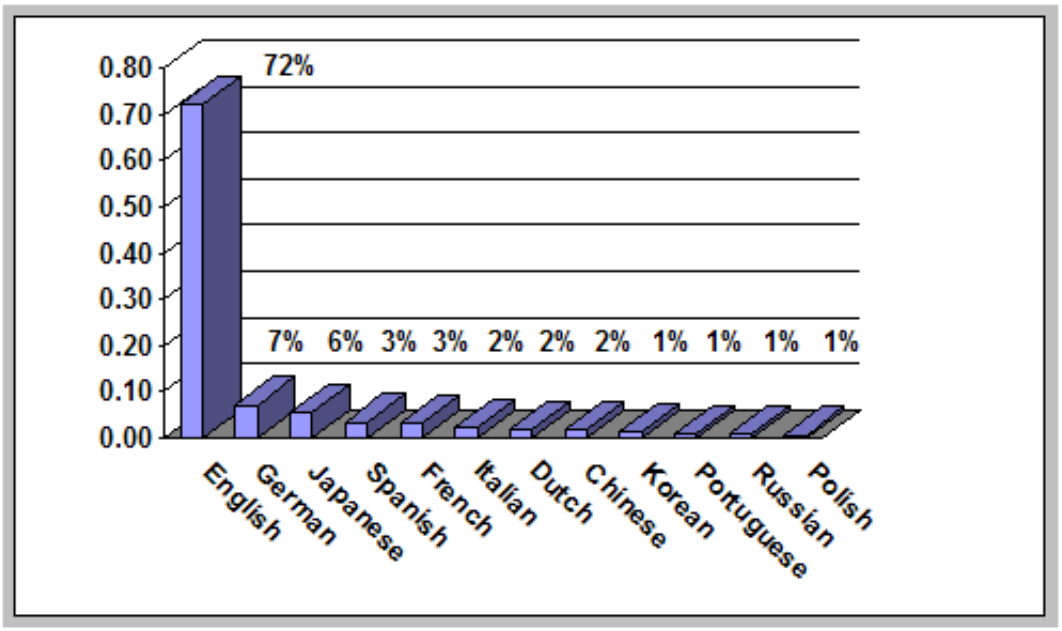


*Рисунок 6: Относительная частота используемых языков, 2002 (процент от общего числа публичных сайтов)*

Это распределение можно сравнить с ситуацией в библиотечных коллекциях. Конечно, веб нельзя сравнивать с одной библиотекой, поскольку ее коллекция создавалась под задачи одной организации. Однако, сравнение веба с агрегатом многих библиотечных коллекций более корректно, т.к. такой подход сглаживает идеографические особенности развития отдельных учреждений. Библиографическая система WorldCat® (объединенный онлайновый каталог OCLC) является крупнейшей в мире и отражает содержание коллекций по всему миру (с доминированием библиотек США). В июле 2001 года каталог WorldCat содержал около 45 миллионов библиографических записей [10]. Из них около 63% ассоциированы с англоязычным контентом. Затем идут немецкий и французский языки – по 6% каждый. Испанский – 4%, китайский, японский, русский и итальянский – по 2% каждый. Все остальные языки не превышали 1%. Такое распределение очень похоже на ситуацию в публичном вебе, как по форме распределение (огромная доля англоязычных ресурсов и быстрый спад в тонкий длинный «хвост» иных языков), так и по относительной частоте использования языков. Это указывает на то, что и веб, и библиотечные коллекции имеют приблизительно одинаковую степень интернационализации с точки зрения языка текстового контента.

## V. Использование метаданных

Библиотеки служат не просто хранилищами информации – они ее еще организуют и индексируют для облегчения поиска и обработки. Говоря о вебе, часто в качестве недостатка называют отсутствие такой организации. Поиск обычно выполняется методом «грубой силы» (например, индексирование ключевых слов), вне контекста и без дополнительных поисковых критериев. По сравнению с первым периодом развития веба произошли улучшения: поисковая система Google, к примеру, использует довольно изощренный алгоритм, который позволяет

ранжировать результаты в соответствии со схемой гиперссылок и популярностью ресурса.

Библиотекари пришли к высокой степени организованности в результате тщательной подготовки и обработки библиографических данных, как например, описательной информации о ресурсах в своих коллекциях. Говоря в общем, такая информация называется метаданными (или «данными о данных»). Уже некоторое время предпринимаются попытки внедрения метаданных в вебе, наиболее заметной из них является инициатива метаданных Дублинского ядра [11]. Был ли достигнут какой-либо прогресс в этом отношении?

Метаданных веб-ресурса обычно размещаются в тэге META, который может быть использован разработчиками для любой информации, которая представляется им релевантной для описания ресурса. Тэг META состоит из двух основных компонентов: NAME (имя), которое идентифицирует определенное поле метаданных (ключевое слово, автор и т.д.) и CONTENT (контент), который содержит значение элемента метаданных, идентифицируемого атрибутом «имя».

На основе данных из всех пяти обзоров «Характеристики веба», можно проанализировать тенденции в использовании метаданных в публичном вебе за последние пять лет. Целью такого анализа было обнаружение любых метаданных в тэге META на публичных сайтах. Было выявлено несколько важных характеристик использования метаданных. Во-первых, очевидно, что такое использование расширяется: наблюдается постоянное увеличение процента публичных веб-сайтов, содержащих метаданные на домашней странице (на которой они используются наиболее часто). Такое же увеличение наблюдается при анализе всех собранных со всех публичных данных веб-страниц, которые содержат метаданные в любой форме.

Необходимо отметить одну особенность: при использовании некоторых HTML-редакторов некоторые META-тэги создаются и заполняются автоматически, как часть шаблона страницы. Вероятно, это объясняет некоторую часть наблюдаемого роста использования этих тэгов на публичных веб-сайтах.

Вторая интересная особенность использования метаданных в вебе заключается в том, что они, очевидно, не становятся более детальными. Если предположить, что один META-тэг эквивалентен одному элементу метаданных (т.е. фрагменту дескриптивной информации о данном ресурсе), то становится очевидным, что в среднем веб-страница включает метаданные о двух-трех элементах. Это трудно назвать широким движением по включению детальных описаний веб-ресурсов в публичном вебе.

*Таблица 1: Использование метаданных на публичных веб-сайтах, 1998 - 2002*

| | 1998 | 1999 | 2000 | 2001 | 2002 |
|---|---|---|---|---|---|
| Число сайтов | 1,457,000 | 2,229,000 | 2,942,000 | 3,119,000 | 3,080,000 |
| % сайтов с метаданными | 70 | 79 | 82 | 85 | 87 |

| | | | | | |
|---|---|---|---|---|---|
| на домашней странице | | | | | |
| Среднее количество тэгов | 2.75 | 2.32 | 2.72 | 2.97 | 3.14 |
| % страниц с метаданными | 45 | 50 | 59 | 62 | 70 |
| Среднее число тэгов | 2.27 | 2.16 | 2.46 | 2.63 | 2.75 |

Среди обескураживающих тенденций в использовании метаданных за последние пять лет – очень низкое использование формальных схем метаданных для описания контента. Например, элементы Дублинского ядра в 1998 году содержались на 0,5% страниц, а в 2002 – всего на 0,7%. Подавляющее большинство метаданных в публичном вебе создаются при разработке сайта и неструктурированны путем имплементации одного из форматов.

## VI. Заключение

На основе пяти ежегодных отчетов проекта OCLC «Характеристики веба» в настоящей работе были проанализированы три ключевых тенденции в эволюции публичного веба. Результаты показывают, что веб развивается весьма быстро и является ценной информационной коллекцией. Однако данные указывают на то, что рост достиг некоторого плато. От года к году публичный веб рос все медленнее, а в последний год даже слегка уменьшился в размере.

Веб изначально позиционируется как глобальный информационный ресурс. Однако анализ показывает, что в публичном вебе доминируют контент-провайдеры из США. Далее, подавляющее большинство контента создано на английском языке. Пока нет признаков того, что в структуре контента веба намечаются сдвиги в сторону более глобального характера.

Наконец, анализ использования метаданных на публичных сайтах показывает, что прогресса в эффективном описании веб-ресурсов для повышения эффективности поиска и обработки, практически не наблюдается. Несмотря на то, что использование мета-тэгов является общим местом, сами метаданные создаются в значительной степени случайным образом. Не наблюдается сколько-нибудь заметной тенденции к использованию формальных схем метаданных для публичных веб-ресурсов.

Рассматривая текущую стадию эволюции веба и обсуждая возможные направления развития, данные настоящей работы указывают на то, что он достиг некоторого водораздела. Период стремительного роста очевидно закончился. Поддержание присутствия в сети становится рутинной и во многих случаях необходимой активностью любых организаций. Однако публичный веб еще далек от полной реализации своего потенциала, как это следует из анализа двух других тенденций в данной статье. В других частях мира не достигнут уровень распространенности веба, характерный для США. Затем, хотя мы и становимся все более умелыми в предоставлении информации в вебе, прогресс в деле

структуризации и организации этой информации весьма невелик.

Прошедшие пять лет стали исключительно убедительными доказательством валидности веба, как «концепции». Надеемся, что следующие пять лет продемонстрируют такой же прогресс в его совершенствовании как с точки зрения удобства пользователей, так и полезности и содержательности его контента.

## Примечания и ссылки

[1] Дополнительную информацию о проекте «Характеристики веба» можно найти на веб сайте <http://wcp.oclc.org/>.

[2] For more information about the Web sampling methodology, please see "A Methodology for Sampling the World Wide Web". Available at <http://www.oclc.org/research/publications/arr/1997/oneill/o'neillar980213.htm>.

[3] Shapiro, C. and H. Varian (1998). Information Rules: A Strategic Guide to the Network Economy, (Harvard Business School Press, Cambridge)

[4] Bergman, M. (2001). "The Deep Web: Surfacing Hidden Value" Journal of Electronic Publishing, Volume 7, Issue 1. Available at <http://www.press.umich.edu/jep/07-01/bergman.html>.

[5] Information on this study, "How Much Information?", along with the study's findings, are available on the project Web site: <http://www.sims.berkeley.edu/research/projects/how-much-info/>.

[6] All ARL-member statistics were obtained at the ARL Statistics and Measurement Program Web site at <http://www.arl.org/stats/index.html>.

[7] Statistics from Gray, M., "Web Growth Summary". Available at <http://www.mit.edu/people/mkgray/net/web-growth-summary.html>.

[8] Kane, M. (2002). "Web Hosting: Only the Strong Survive" ZDNet News. Available at <http://zdnet.com.com/2100-1105-938347.html>.

[9] Mariano, G. (2002). "The Incredible Shrinking Internet" ZDNet UK News. Available at <http://news.zdnet.co.uk/story/0,,t269-s2101890,00.html>.

[10] WorldCat statistics were obtained from the OCLC Annual Report 2000/2001. The report is available at <http://www.oclc.org/about/annualreport/2001.pdf>.

[11] For more information about the Dublin Core Metadata Initiative, please visit the DCMI Web site at <http://dublincore.org/>.

© *E.T.O'Neill, B.F. Lavoie, R. Bennett, 2003*

*(c) Перевод С.А. Арнаутова, 2003*