

# HG2052

## Language, Technology and the Internet

### The World Wide Web and HTML

Francis Bond

Division of Linguistics and Multilingual Studies

<http://www3.ntu.edu.sg/home/fcbond/>

[bond@ieee.org](mailto:bond@ieee.org)

Lecture 6

# Revision of Collaboration and Wikis

---

- Version Control Systems
- Wikipedia
- Licensing and Ownership

# Version Control Systems

---

- Versioning file systems
  - every time a file is opened, a new copy is stored
- CVS, Subversion, Git
  - changes to a collection of files are tracked
  - simultaneous changes are merged
- Revision Tracking
  - Revisions are stored within a file
- Authorship in shared writing

# Wikipedia

---

- The core aim of the Wikimedia Foundation, is to get a free encyclopedia to every single person on the planet. (Jimmy Wales)
- Wikipedia makes it easy to share your knowledge  
people like to do this
- Most edits are done by insiders!
- Most content is added by outsiders!
- Content comparable to Britannica

# The five pillars of Wikipedia

---

1. Wikipedia is an online encyclopedia
2. Wikipedia has a neutral point of view.
3. Wikipedia is free content
4. Wikipedians should interact in a respectful and civil manner
5. Wikipedia does not have firm rules

# Licenses and Ownership

---

- Copyright
- Copyleft
- Creative Commons

# What is a good article?

---

1. Well-written
2. Factually accurate and verifiable
3. Broad in its coverage
4. Neutral
5. Stable
6. Illustrated, if possible, by images

---

# The World Wide Web and HTML



# Overview

---

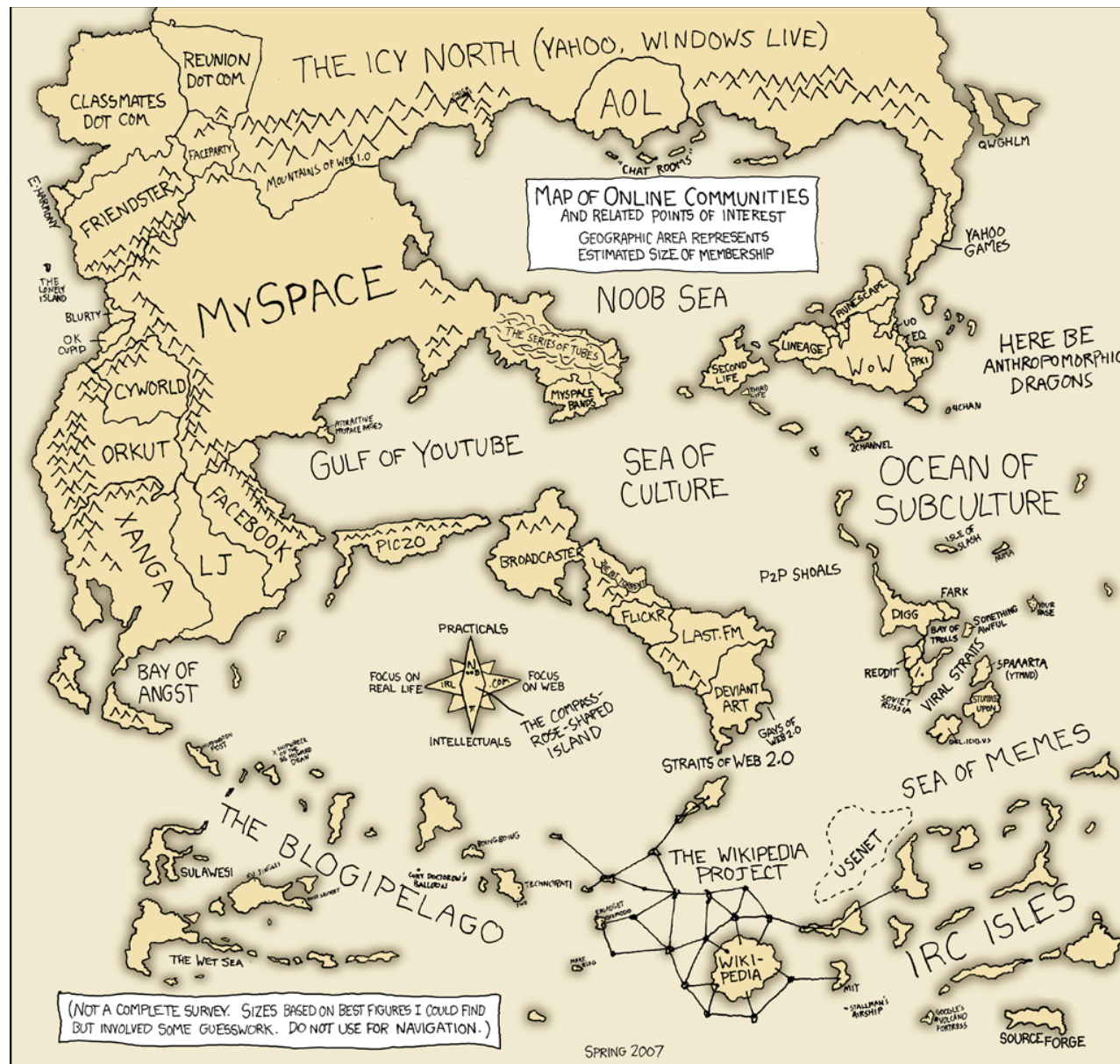
- The Internet
- The structure of Markup
- The structure of the Web
- The future of the Web
- Linguistic features of the web

# The Internet

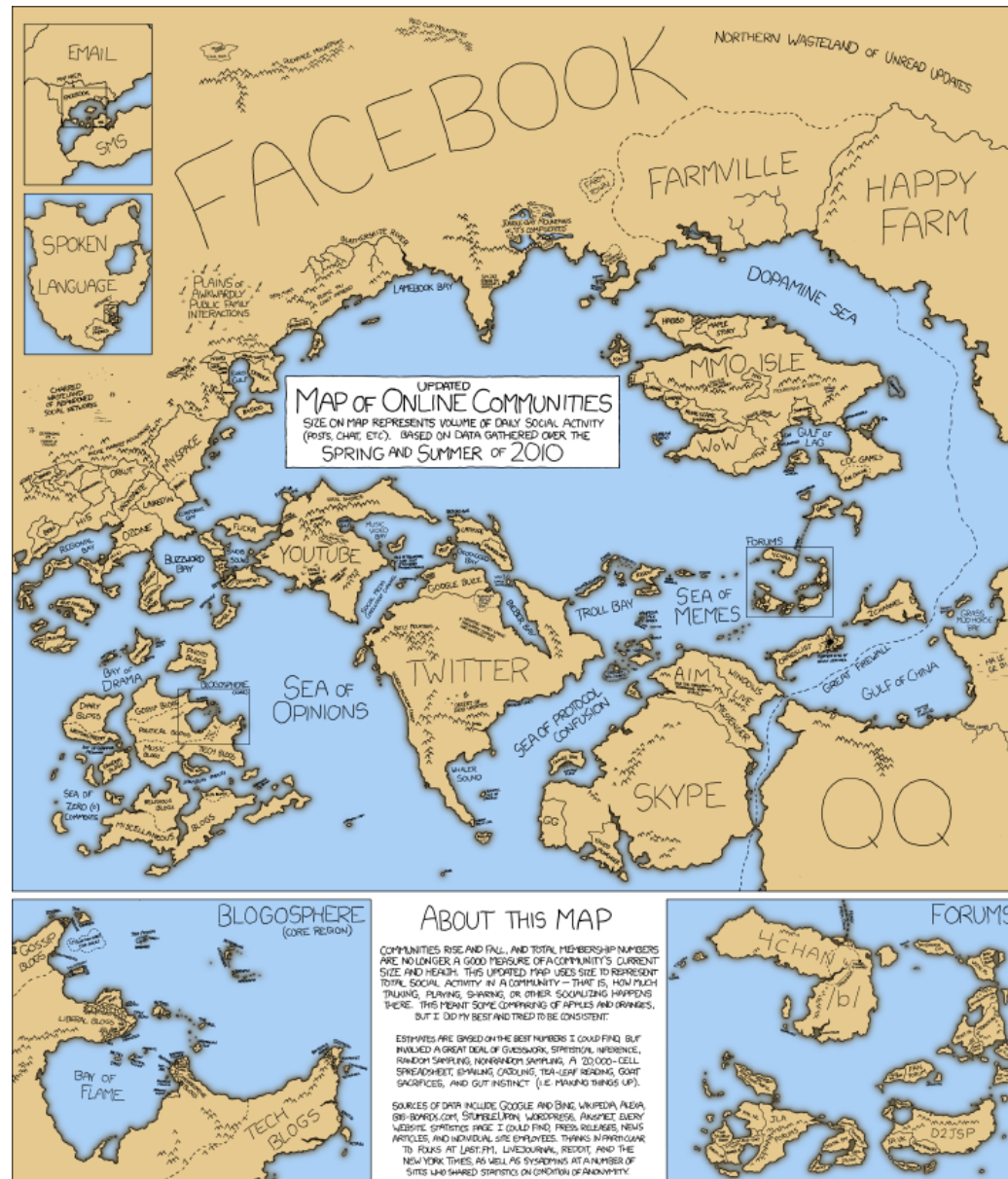
---

- global system of interconnected computer networks that use the standard Internet Protocol Suite (TCP/IP)
- Carries several services
  - HTTP (Hyper Text Transfer Protocol) — The Web
  - Email
  - VoIP (Voice over IP) — Telephony/Skype
  - FTP, ...(File Transfer)
  - Streaming Media — music, video
  - Instant Messaging

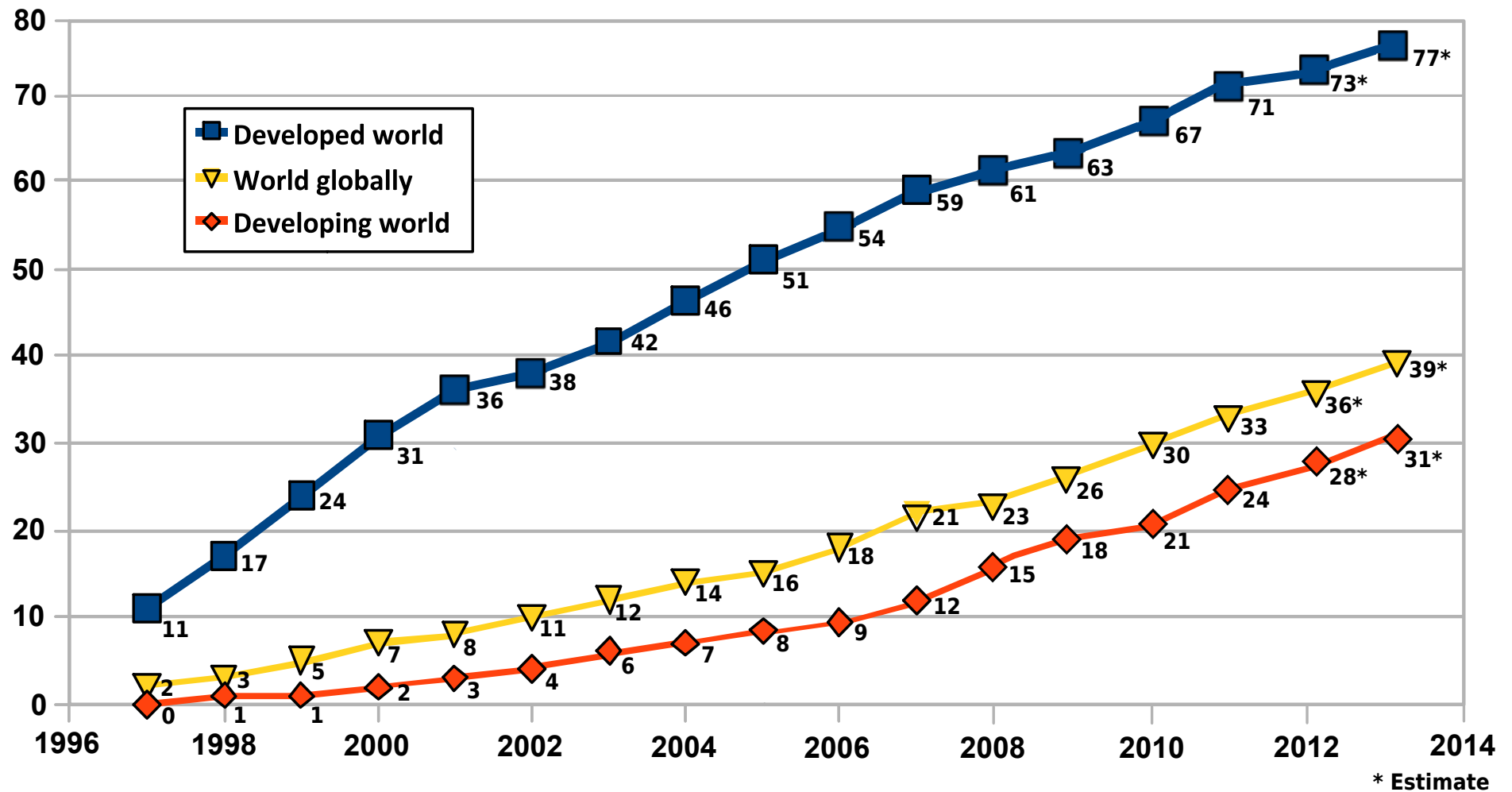
## Map of online communities (2007)



# Map of online communities (2010)



# Growth of the Internet



---

# Markup formatting information

# Why Markup?

---

- Reduce Ambiguity
  - Need to make meaning explicit
- Traditionally this is done by annotating text in some way

# Markup Languages

---

- Annotation on how to print is called **markup**
  - underlining to indicate boldface
  - special symbols for passages to be omitted
  - special symbols for printed in a particular font
- This existed before computers
  - Editors would markup hand-written manuscripts
  - ...and pass them to type setters
  - ...who would prepare the manuscript for printing



# Printers' Markup

## Style of Type

Wf

Wrong font (size or style of type)

lc

lower case letter

lc

Set in LOWER CASE

C

capital letter

Caps

SET IN capitals

C+lc

Set in lower case with INITIAL CAPITALS

sc

SET IN small CAPITALS

C+sc

SET IN SMALL CAPITALS with initial capitals

Eq #

Equalize space between words

the/3

## Insertion and Deletion

Caret (insert marginal addition)

g

Delete (take it out)

e

Correct letter or word marked

stat

Let it stand (all matter above dots)

## Paragraphing

¶

Begin a paragraph

# Early Computer Markup (troff)

---

## Headline

and some text

```
.ps 12                % point size 12
.ft B                 % font type Bold
Headline
.ps 10                % point size 10
.ft R                 % font type Roman
and some text.
```

- Marked up with troff
- Postscript and PDF (Portable Document Format) are similar

# Visual Markup vs Logical Markup

---

## ➤ **Visual Markup** (Presentational)

- What you see is what you get (**WYSIWYG**)
- Equivalent of printers' markup
- Shows what things look like

## ➤ **Logical Markup** (Structural)

- Shows the structure and meaning
- Can be mapped to visual markup
- Less flexible than visual markup
- More adaptable (and reusable)

# Standard Generalize Markup Language: SGML

---

- ISO standard based on IBM's GML
- Attempt to make markup independent of processor
  - Important for archiving information
- Emphasis on logical markup
- Popularized the use of `<tag></tag>` notation
  - and entities `&lt;`; `&gt;`; when you need an `<>`
- Split the document into: Declaration, Prolog, Documentation

# Hyper Text Markup Language: HTML

---

- Markup Language for web pages
- An extension of SGML
- Combines logical and visual markup
- Also allows hyperlinks (linking and anchoring)
- Created by Tim Berners-Lee at CERN (1989)
  - to make physics papers and documentation more accessible

# HTML example

---

## Headline

and some text

### ➤ Logical

```
<h1>Headline</h1>
```

```
<p>and some text
```

### ➤ Visual

```
<font size="3"><b>Headline</b></font>
```

```
<br>and some text
```

# Logical allows various styles

---

Headline

and some text

```
<style>
H1 {
    font-size:24px;
    color:blue;
    margin-top:10px;
    margin-bottom:15px;
}
</style>
```

- This can be done using CSS (Cascading Style Sheets)
- Separate Logical and Visual Structure

# Benefits of Logical Tags

---

- Can transform things easily
  - No bold for Japanese and Chinese (just use size)
  - Can adapt to other modalities (speech)
- Logical form useful for other tasks
  - Summarization
    - \* Just show `<h1>` ... `<h3>`
  - Translation
    - \* Headers are noun phrases, not sentences
- Robustness: you can read the source directly



## But still there is ambiguity!

---

- Tags on one site may not mean the same thing on another site
- Huge amount of information
  - Looking for **Eric Miller** may get the wrong one!
  - Looking for **NTU** gets
    - \* Nanyang Technological University
    - \* National Taxpayers Union
    - \* National Taiwan University
- What can we do?  
Semantic Web (week 10)

# Hypertext

---

- HTML crucially adds [hyperlinks](#)
  - these extend text in a new way
  - references that you can immediately access
- `<href="http://somewhere.on.the.web">link me</a>`
- ``
- Immediately accessible references are qualitatively different

# HTML example

---

```
<!doctype html>
<html>
  <head>
    <title>Hello HTML</title>
  </head>
  <body>
    <p>Hello World!</p>
    <p>Oh well, <span lang="fr">c'est la vie</span>,
      as they say in France.</p>
    <abbr id="anId" class="jargon" style="color:blue;"
      title="Hypertext Markup Language">HTML</abbr>
  </body>
</html>
```

# How should you hyperlink?

---

- Pick a page
  - This course page
  - LMS research page
  - Wiki front page
  - Your choice
- Discuss whether you think there are enough links or too many or not enough? And are they linking to the best targets?
- You may wish to look at the *Wikipedia:Manual of Style/Linking*  
<[https://en.wikipedia.org/wiki/Wikipedia:Manual\\_of\\_Style/Linking](https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Linking)>

# The Structure of the Web

---

- 550 billion documents on the Web (2001)  
mostly in the invisible Web, or deep Web
- 11.5 billion indexable web pages (2005)
- 25.21 billion indexable web pages (2009)
- 109.5 million websites (2009)

# The Deep Web

---

**Dynamic content** dynamic pages which are returned in response to a submitted query or accessed only through a form

**Unlinked content** pages which are not linked to by other pages (but clicking links them)

**Private Web** sites that require registration and login ([Edventure](#), [NTULearn](#))

**Contextual Web** pages with content varying for different access contexts (e.g., ranges of client IP addresses or previous navigation sequence).

**Limited access content** sites that limit access to their pages in a technical way (e.g., using the Robots Exclusion Standard)

**Scripted content** pages that are only accessible through links produced by JavaScript as well as content dynamically downloaded from Web servers via Flash or Ajax solutions.

---

**Non-HTML/text content** textual content encoded in multimedia (image or video) files or specific file formats not handled by search engines.

These pages all include data that search engines cannot find!

# robots.txt

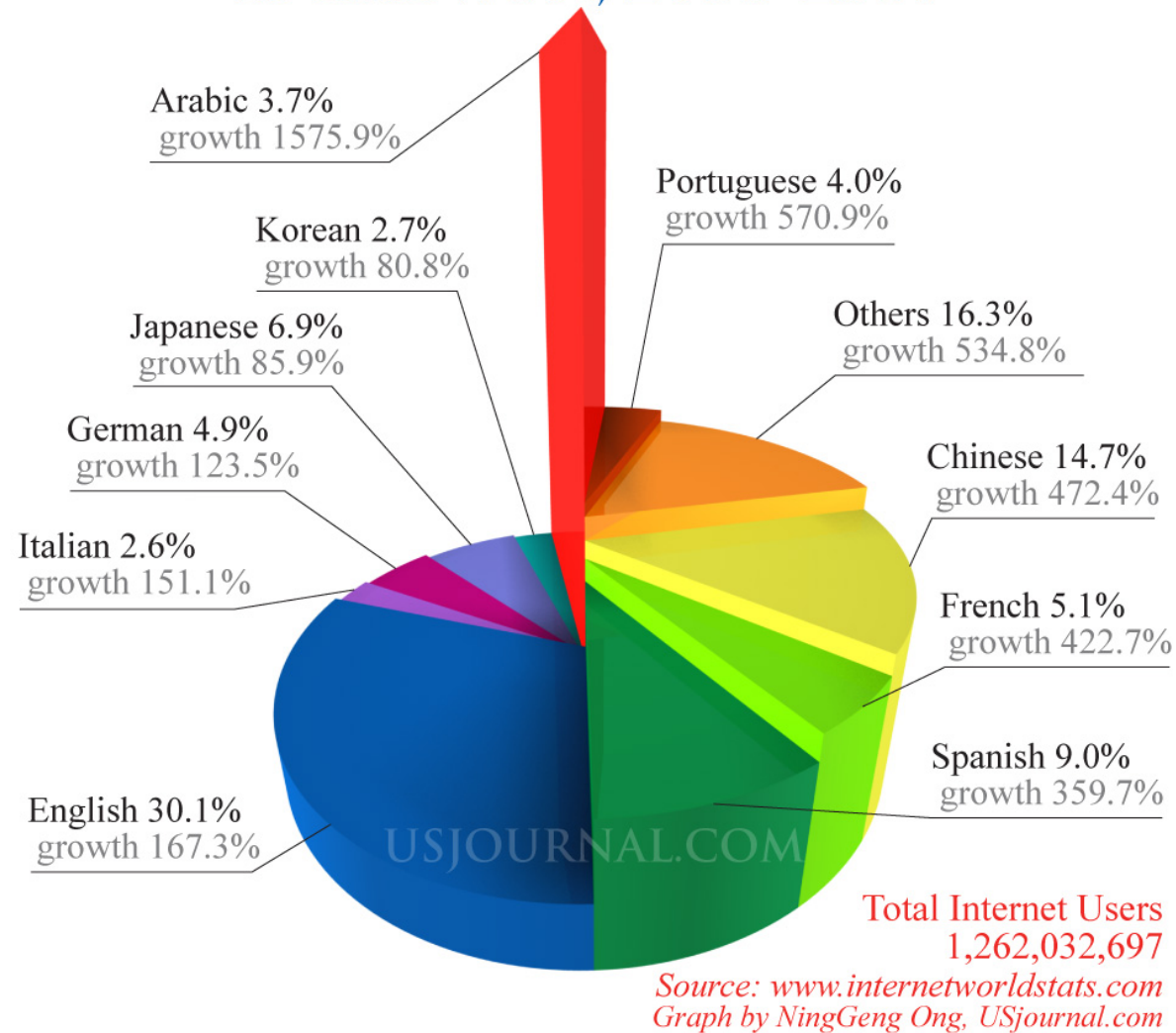
---

- A **Robot** (Web Crawler, or Spiders) is a program that automatically traverses the Web's hypertext structure by retrieving a document, and recursively retrieving all documents that are referenced. Robots are used for:
  - Indexing and *What's New* monitoring
  - HTML and Link validation
  - Mirroring and back up
- A website can explicitly tell robots where they can and cannot go
  - Compliance is voluntary, but followed by most robots
- You can **Allow** and **Disallow** whole directories, or individual pages
- You can **Allow** and **Disallow** individual user-agents (such as Google)

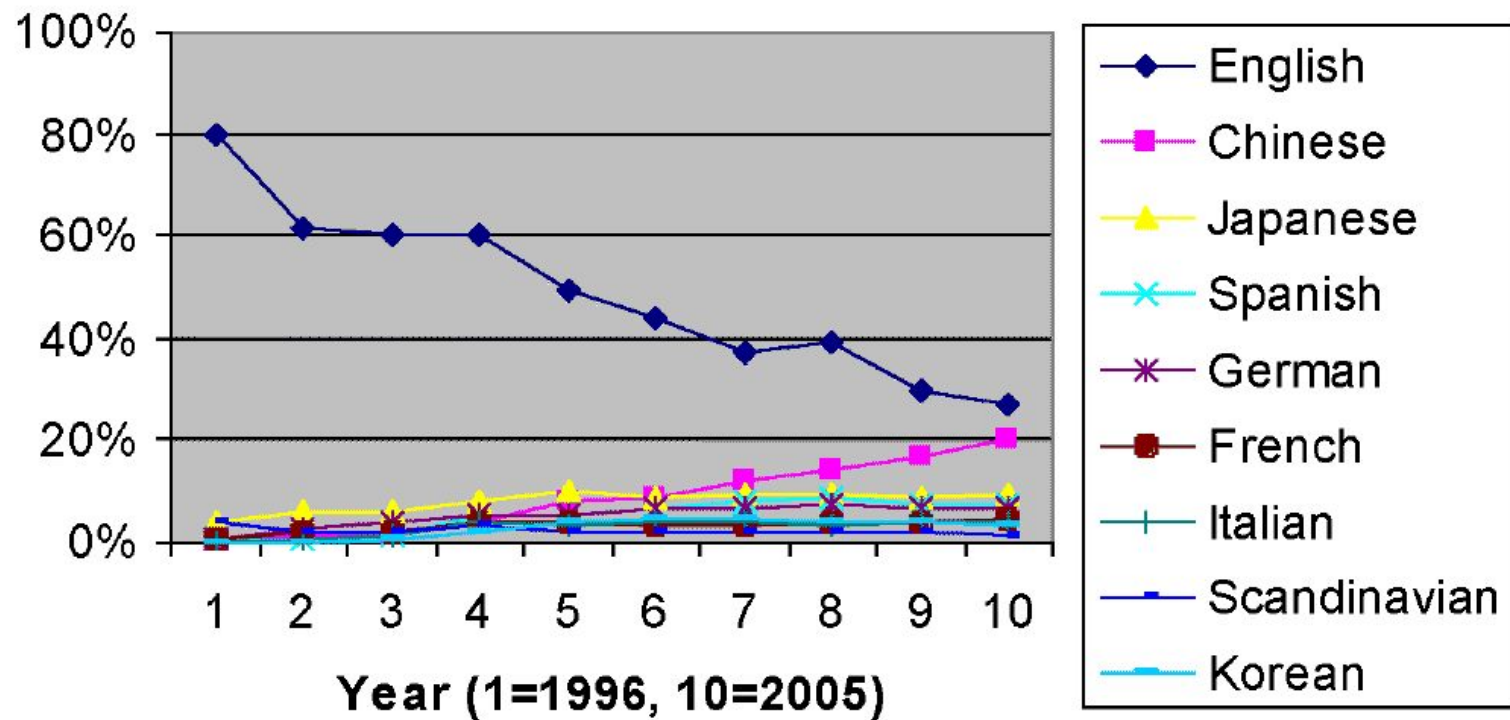


# The Internet and Language Diversity

## INTERNET USAGE BY LANGUAGE 2007 & GROWTH, 2000-2007

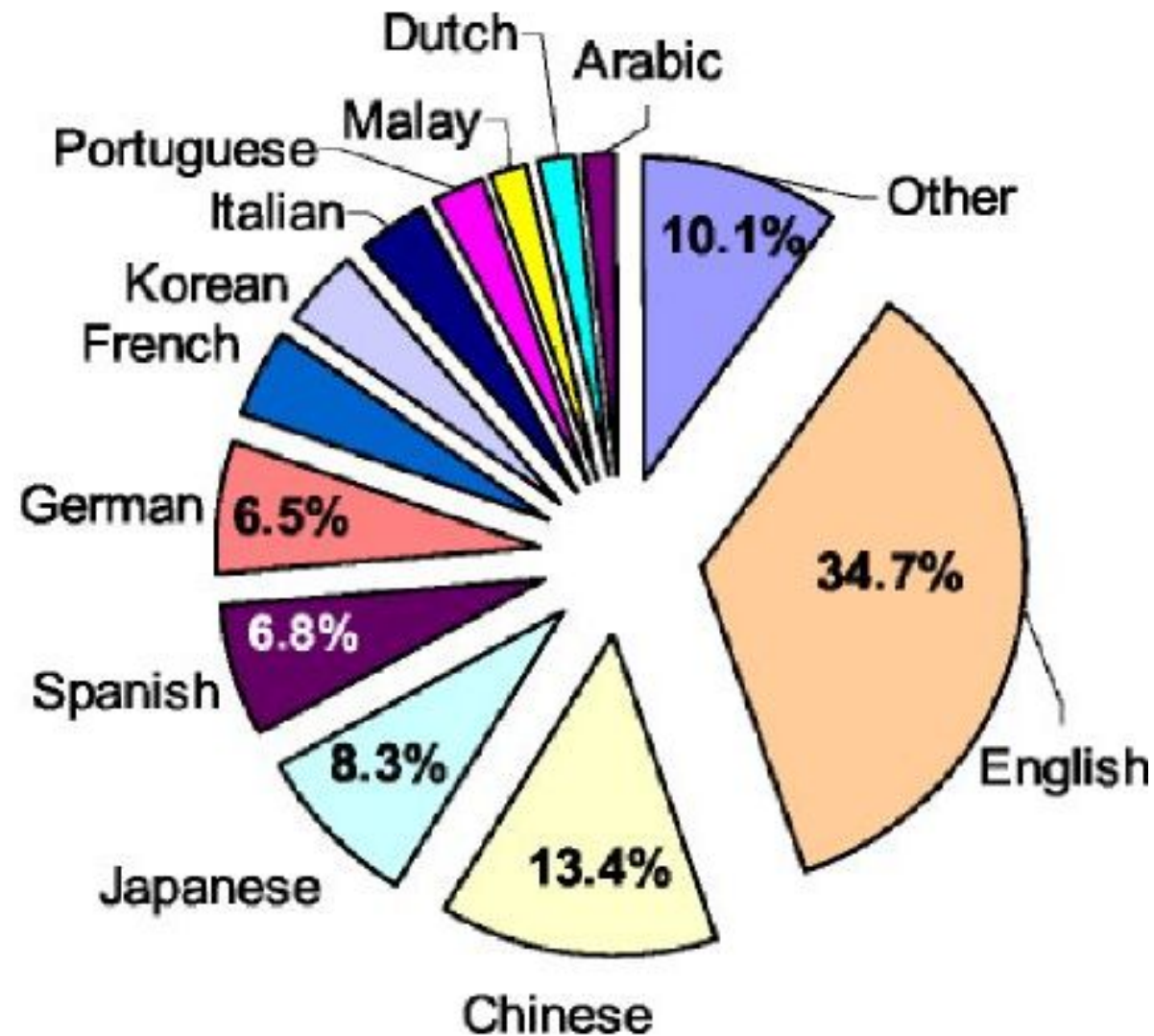


# Distribution of languages among Internet users

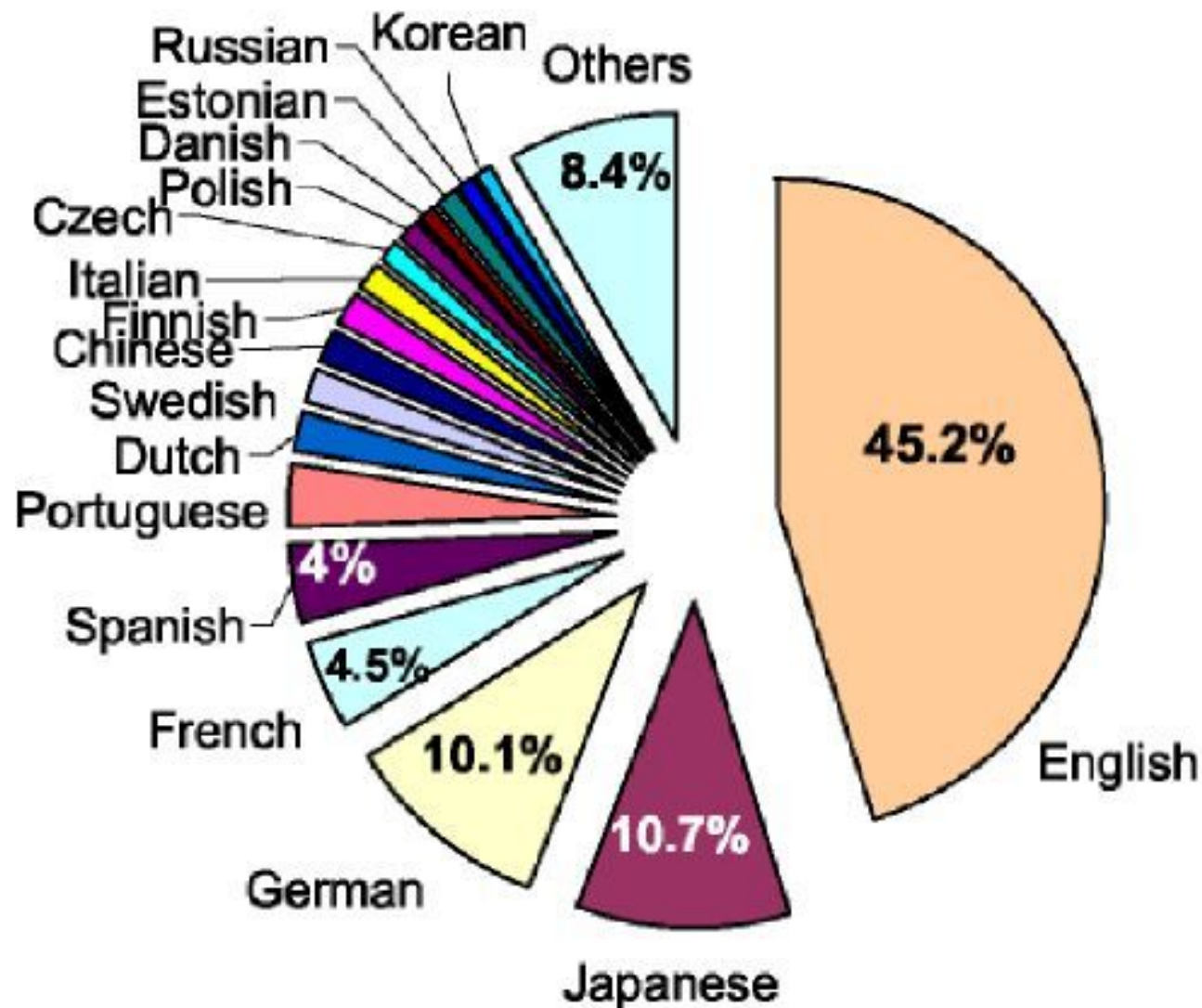


## Internet users by language, February 2005

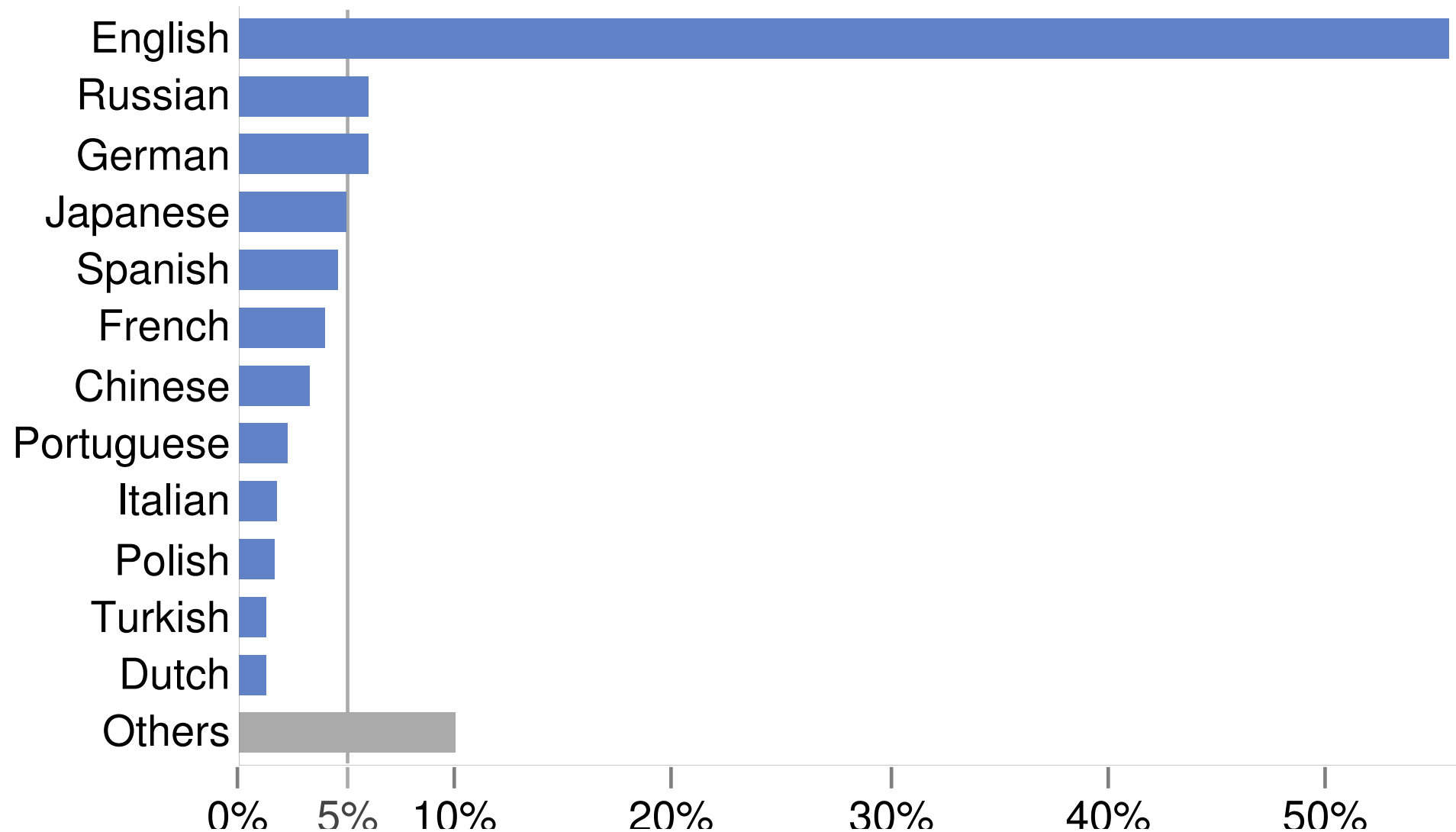
---



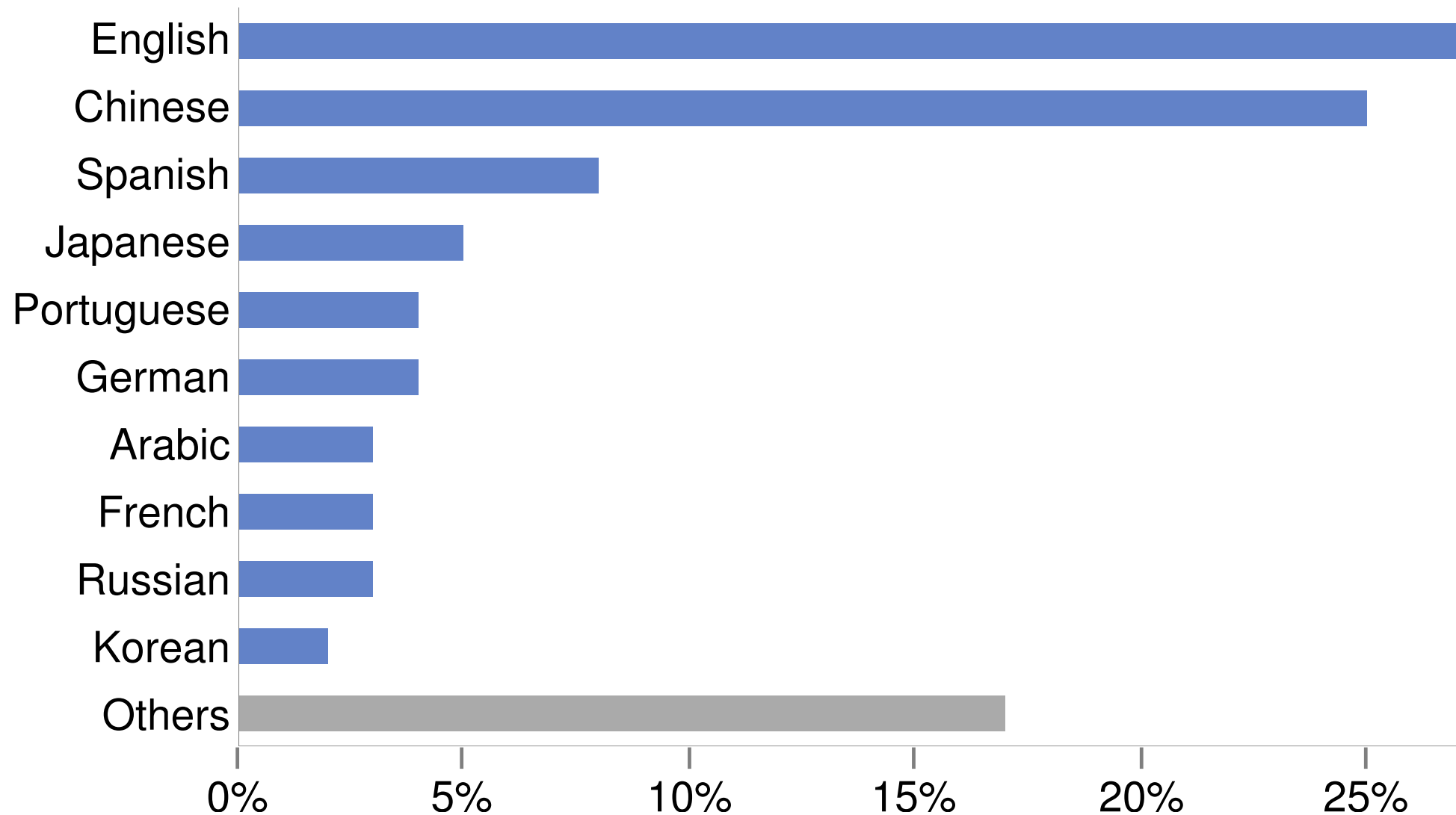
## Language of e-commerce, February 2005



## Percentage of Web sites by language (2014)

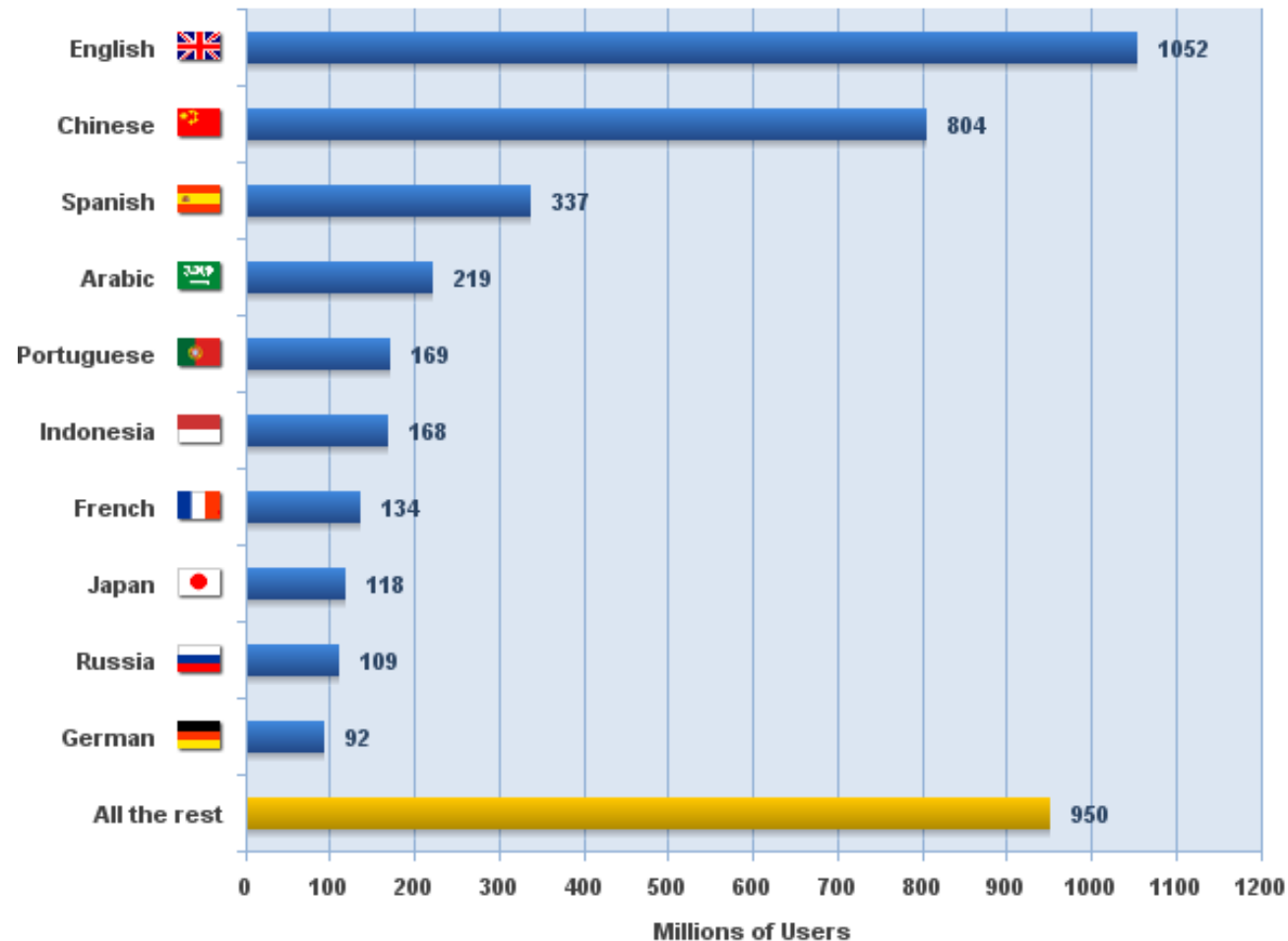


## Percentage of Web users by language (2014)



# Gradually Changing

**Top Ten Languages in the Internet  
in Millions of users - December 2017**



Source: Internet World Stats - [www.internetworldstats.com/stats7.htm](http://www.internetworldstats.com/stats7.htm)  
Estimated total Internet users are 4,156,932,140 in December 31, 2017  
Copyright © 2018, Miniwatts Marketing Group

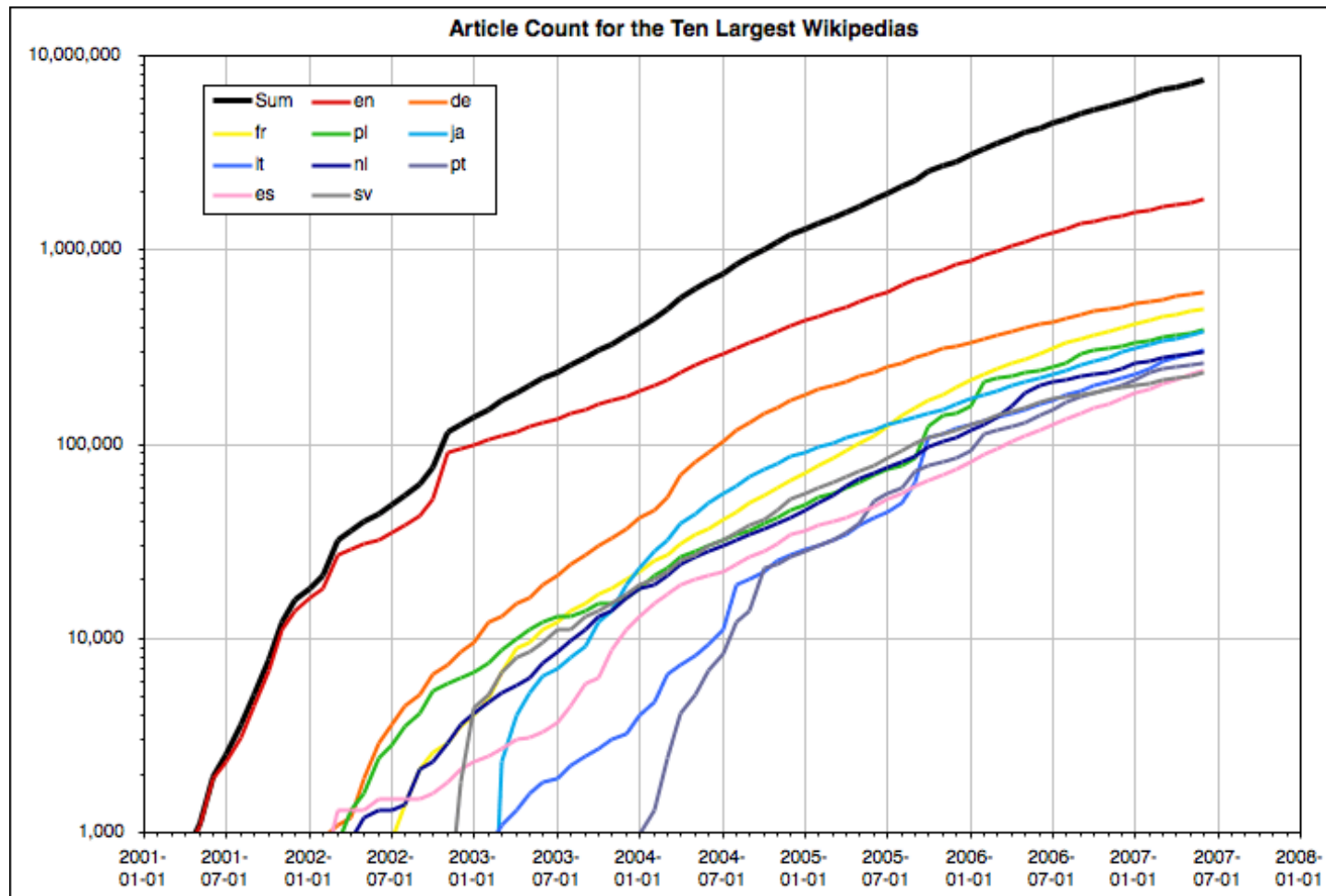
# The Internet and Language Diversity

---

- Major languages will survive (not just English)
- **Sarnoff's Law:** the value of a broadcast network is proportional to the number of viewers ( $n$ )
- **Metcalf's Law:** the value of a telecommunications network is proportional to the square of the number of connected users of the system ( $n^2$ )
  - ⇒ languages with more pages will become even more valuable
- Minor languages probably won't survive



# Top ten Wikipedias



See also [http://meta.wikimedia.org/wiki/List\\_of\\_Wikipedias](http://meta.wikimedia.org/wiki/List_of_Wikipedias)  
Wikipedias in 272 languages: only 96 with more than 10,000 pages

# The next 5,000 days of the Web

---

- Kevin Kelly on the next 5,000 days of the web (20min)
- [http://www.ted.com/talks/lang/eng/kevin\\_kelly\\_on\\_the\\_next\\_5\\_000\\_days\\_of\\_the\\_web.html](http://www.ted.com/talks/lang/eng/kevin_kelly_on_the_next_5_000_days_of_the_web.html)
- The impossible has become possible
- The web is a single machine
  - Embodiment
  - Re-structuring
  - Co-dependence

# Linguistic features of the web

---

- Much/most text is just the same
- Un-edited
- Accessible in great volume (and many languages)
- Editable — Wikis, comments, tweets
- Multi-media

# Conclusion

---

- The web is changing what humanity can do with language
- It is not clear if it is changing what individual humans do
- Make sure you go through the wikipedia tutorial

## References

---

- Crystal, D. (2011). *Internet Linguistics: a student guide*. Routledge
- Peter Gerrand (2007) Estimating linguistic diversity on the Internet: A taxonomy to avoid pitfalls and paradoxes. *Journal of Computer-Mediated Communication*, 12(4), article 8. <http://jcmc.indiana.edu/vol12/issue4/gerrand.html>
- Global Reach. (2006). Global Internet Statistics (by Language). Retrieved October 11, 2006 from <http://www.global-reach.biz/globstats/index.php3>