

Introduction to the Web

MPRI 2.26.2: Web Data Management

Antoine Amarilli

Friday, December 7th



The old days

1969 ARPANET (ancestor of the Internet)

1974 TCP

1990 The World Wide Web, HTTP, HTML

1994 Yahoo! was founded

1995 Amazon.com, Ebay, AltaVista are founded

1998 Google are founded

2001 Wikipedia is created

- Around **330 million** domains, including 130 million in `.com`¹
- **54%** of content in **English** and **4%** in French²
- **48%** of people have Internet access, and **71%** of ages 15–24³
- Google knows over one **trillion** (10^{12}) of unique URLs⁴
 - The **same content** can live in many different URLs
 - Parts of the Web are not indexable: the **hidden Web** or **deep Web**

¹<https://www.businesswire.com/news/home/20170914006386/en/Internet-Grows-331.9-Million-Domain-Registrations-Quarter>

²https://w3techs.com/technologies/overview/content_language/all

³<https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2017.pdf>

⁴<https://googleblog.blogspot.fr/2008/07/we-knew-web-was-big.html>

Table of Contents

Introduction

Web browsers

Other Web clients

URLs

Historical web browsers

Mosaic First common graphical browser, 1993–1997

- From 80% in 1994 to < 10% in 1996

Historical web browsers

Mosaic First common graphical browser, 1993–1997

- From 80% in 1994 to < 10% in 1996

Netscape Released in 1994, based on Mosaic

- From 80% in 1996 to < 10% in 2001

Historical web browsers

Mosaic First common graphical browser, 1993–1997

- From 80% in 1994 to < 10% in 1996

Netscape Released in 1994, based on Mosaic

- From 80% in 1996 to < 10% in 2001

Internet Explorer Released in 1995

- Provided with Windows 95
- IE 6 released in 2001 and reaches 80% market share
- Leads to an antitrust lawsuit in the USA, 1998–2001

Historical web browsers

Mosaic First common graphical browser, **1993–1997**

- From **80%** in 1994 to **< 10%** in 1996

Netscape Released in **1994**, based on Mosaic

- From **80%** in 1996 to **< 10%** in 2001

Internet Explorer Released in **1995**

- Provided with **Windows 95**
- IE 6 released in 2001 and reaches **80%** market share
- Leads to an **antitrust** lawsuit in the USA, 1998–2001

Firefox Released in **2002** from Netscape (open-sourced in 1998)

- **Free** and **open-source**
- Popularized **tabbed browsing**
- Attacked IE 6's **monopoly**

Current Web browsers

IE IE 7 released in 2006, replaced by **Microsoft Edge**

Firefox Still **actively developed**

Safari Released in 2003, default Web browser on **Mac OS X**

Opera Released in **1996**, initially **commercial** (until 2000) then **ad-supported** (until 2005), now **free** but **proprietary**.

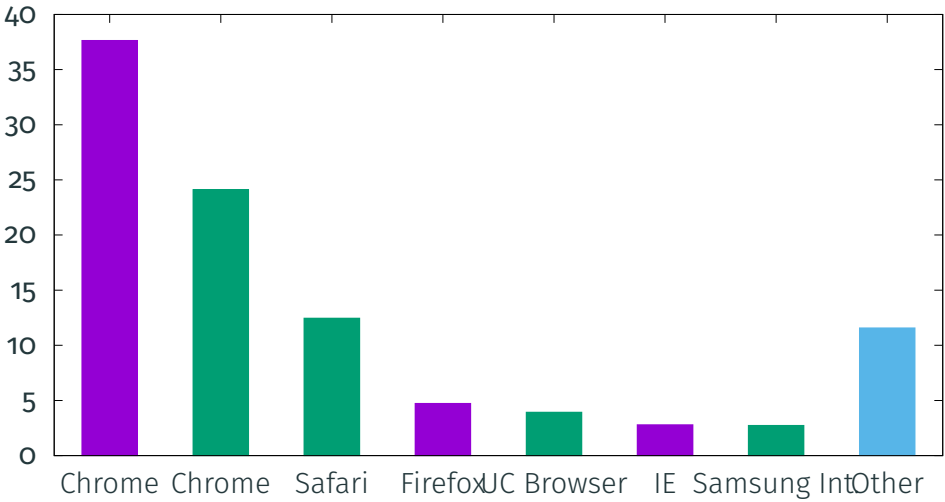
Chrome Released in **2008** by **Google**, with an **open-source version** (Chromium)

Mobile **55%** market share according to `gs.statcounter.com`

- **Safari** (iOS), **Android browser** or **Chrome** (Android)
- Firefox Mobile, Blackberry, Opera, UC Browser...

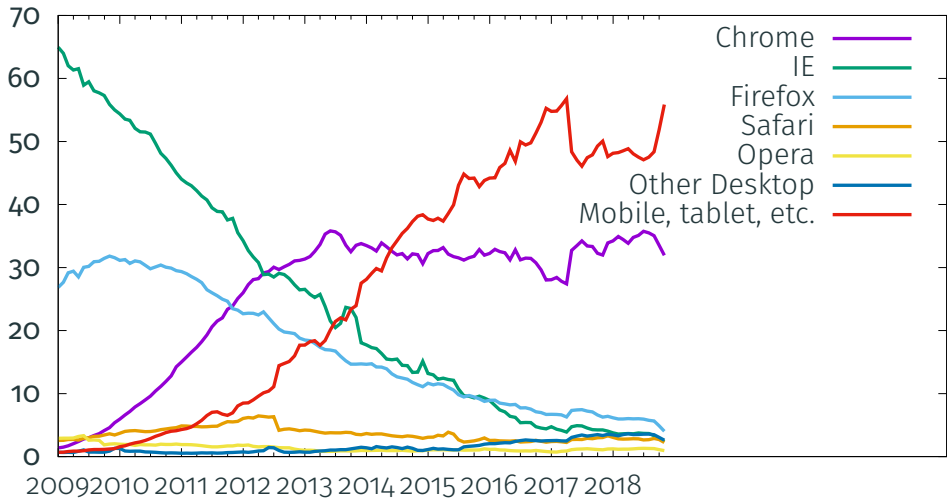
To check rendering on old browsers, use `browsershots.org`

Recent market share



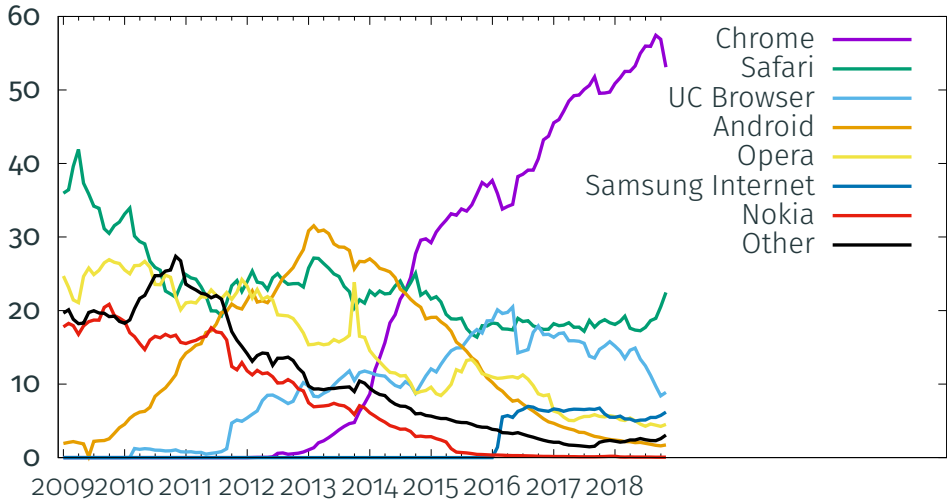
Source: gs.statcounter.com (November 2018)

Evolution



Source: gs.statcounter.com

Evolution (mobile)



Source: gs.statcounter.com

Rendering engine

Firefox Gecko, and (work-in-progress) Servo, using Rust

Safari WebKit engine

Chrome Blink (fork of Webkit, in April 2013)

IE Originally Trident, then EdgeHTML, soon Chromium⁵

Opera Originally Presto, then Blink

Others Dillo, KHTML, and other old/minimalistic engines

⁵<https://www.windowscentral.com/>

Summary and perspectives

- Webkit/Blink and Chrome/Chromium are **dominant**
- Only serious challenger: **Firefox** (but around **5%** globally)
- Blink is **open-source** but **controlled by Google**
- Different **browsers** using this rendering engine
 - Some **minimalistic**, e.g., `uzbl`
 - Some **variants** of existing browsers, e.g., Pale Moon, Basilisk (Firefox forks)
 - Some **new browsers** using Blink: Vivaldi, Brave

New topics about Web browsers

- Ecosystem of **extensions**: Chrome Web Store, Mozilla Store
 - Firefox: Mandatory **signing**, and **backward-incompatible** changes

⁶<https://www.statista.com/topics/3201/ad-blocking/>

New topics about Web browsers

- Ecosystem of **extensions**: Chrome Web Store, Mozilla Store
 - Firefox: Mandatory **signing**, and **backward-incompatible** changes
- **Ad blocking** (27% of users⁶), counter*-measures
 - **uBlock Origin** extension, based on **Easylist**
`https://easylist.to/easylist/easylist.txt`
 - More generally, **JavaScript blockers**, e.g., uMatrix, NoScript, etc.

⁶<https://www.statista.com/topics/3201/ad-blocking/>

New topics about Web browsers

- Ecosystem of **extensions**: Chrome Web Store, Mozilla Store
 - Firefox: Mandatory **signing**, and **backward-incompatible** changes
- **Ad blocking** (27% of users⁶), counter*-measures
 - **uBlock Origin** extension, based on **Easylist**
`https://easylist.to/easylist/easylist.txt`
 - More generally, **JavaScript blockers**, e.g., uMatrix, NoScript, etc.
- Filtering out **bots**: robots exclusion standard, CAPTCHAs
 - **reCAPTCHA**: now volunteer work for Google

⁶<https://www.statista.com/topics/3201/ad-blocking/>

New topics about Web browsers

- Ecosystem of **extensions**: Chrome Web Store, Mozilla Store
 - Firefox: Mandatory **signing**, and **backward-incompatible** changes
- **Ad blocking** (27% of users⁶), counter*-measures
 - **uBlock Origin** extension, based on **Easylist**
`https://easylist.to/easylist/easylist.txt`
 - More generally, **JavaScript blockers**, e.g., uMatrix, NoScript, etc.
- Filtering out **bots**: robots exclusion standard, CAPTCHAs
 - **reCAPTCHA**: now volunteer work for Google
- **Security**: site isolation, one process/site (in Chromium)

⁶<https://www.statista.com/topics/3201/ad-blocking/>

New topics about Web browsers

- Ecosystem of **extensions**: Chrome Web Store, Mozilla Store
 - Firefox: Mandatory **signing**, and **backward-incompatible** changes
- **Ad blocking** (27% of users⁶), counter*-measures
 - **uBlock Origin** extension, based on **Easylist**
`https://easylist.to/easylist/easylist.txt`
 - More generally, **JavaScript blockers**, e.g., uMatrix, NoScript, etc.
- Filtering out **bots**: robots exclusion standard, CAPTCHAs
 - **reCAPTCHA**: now volunteer work for Google
- **Security**: site isolation, one process/site (in Chromium)
- Web browser **fingerprinting** `https://panopticlick.eff.org/`

⁶`https://www.statista.com/topics/3201/ad-blocking/`

New topics about Web browsers

- Ecosystem of **extensions**: Chrome Web Store, Mozilla Store
 - Firefox: Mandatory **signing**, and **backward-incompatible** changes
- **Ad blocking** (27% of users⁶), counter*-measures
 - **uBlock Origin** extension, based on **Easylist**
`https://easylist.to/easylist/easylist.txt`
 - More generally, **JavaScript blockers**, e.g., uMatrix, NoScript, etc.
- Filtering out **bots**: robots exclusion standard, CAPTCHAs
 - **reCAPTCHA**: now volunteer work for Google
- **Security**: site isolation, one process/site (in Chromium)
- Web browser **fingerprinting** `https://panopticklick.eff.org/`
- In-browser cryptocurrency mining: **CoinHive**
 - around **0.01 \$/day**, vs about **0.10 \$/day** of electricity

⁶<https://www.statista.com/topics/3201/ad-blocking/>

New topics about Web browsers

- Ecosystem of **extensions**: Chrome Web Store, Mozilla Store
 - Firefox: Mandatory **signing**, and **backward-incompatible** changes
- **Ad blocking** (27% of users⁶), counter*-measures
 - **uBlock Origin** extension, based on **Easylist**
`https://easylist.to/easylist/easylist.txt`
 - More generally, **JavaScript blockers**, e.g., uMatrix, NoScript, etc.
- Filtering out **bots**: robots exclusion standard, CAPTCHAs
 - **reCAPTCHA**: now volunteer work for Google
- **Security**: site isolation, one process/site (in Chromium)
- Web browser **fingerprinting** `https://panopticklick.eff.org/`
- In-browser cryptocurrency mining: **CoinHive**
 - around **0.01 \$/day**, vs about **0.10 \$/day** of electricity
- **Tor** and **Tor hidden services**

⁶<https://www.statista.com/topics/3201/ad-blocking/>

Table of Contents

Introduction

Web browsers

Other Web clients

URLs

Textual Web browsers

```
sampi:~$ w3m 'http://en.wikipedia.org' (14:24:59)
Télécom ParisTech

From Wikipedia, the free encyclopedia
(Redirected from Telecom ParisTech)
Jump to: navigation, search
"ENST" redirects here. For the airport with this ICAO airport code, see Sandnessjøen Airport, Stokka.

Crystal Clear This article may need to be rewritten entirely to comply with Wikipedia's quality
app kedit.svg standards. You can help. The discussion page may contain suggestions. (May 2009)

Coordinates: 48°49′35″N 2°20′47″E﻿ / ﻿48.82639°N 2.34639°E﻿ / 48.82639; 2.34639

Télécom ParisTech

Logo telecomparisTech.png

Motto L'École au coeur de la Société de l'Information

Established 1878

Type French Grande École

President Yves Poilane

Admin. staff 340 (2006)

Students 1249 (2006)

Location Paris, France

Campus Paris, Sophia Antipolis

⏏ ↑ ↓ Viewing <Télécom ParisTech - Wikipedia, the free encyclopedia>
```

- **lynx** (still maintained), w3m, elinks
- Also: **screen readers** for visually impaired users

Many **automated programs** on the Web:

- Search engine **crawlers**: see Pierre's class
- **RSS readers** and aggregators
- **Email harvesters** (spammers)
- **API consumers**

Table of Contents

Introduction

Web browsers

Other Web clients

URLs

- Anatomy of a URL:

`https://en.wikipedia.org/wiki/Telecom_ParisTech#History`
protocol machine path fragment

- Uniform Resource Locator
- Identifies a **resource** on the Web

- Course material inspired by course notes by Pierre Senellart