What is the invisible web? A crawler perspective

Natalia Arroyo

Internet Lab

CINDOC-CSIC

Narroyo@cindoc.csic.es

Invisible web and crawlers

- **Invisible web** is the part of the Internet that can not be indexed by search engines
- **Crawlers** are programs that automatically fetch Web pages
- What is invisible web from a crawler perspective? Is it the same than for search engines?
- **Purpose**: to introduce a new perspective in the concept of invisible web

Methodology

- Commercial crawlers
  - Astra SiteManager 2.0
  - COAST Web Master 7.0
  - Microsoft Site Analyst 2.0
  - Microsoft Content Analyzer 3.0
  - Xenu 1.2f
- Academic crawler
  - SocSciBot 1.8.108
- Institutional web sites from the European academic webspace

Classification

Search engines	Crawlers
Opaque web	
Private web	Private web*
Proprietary web	Proprietary web
Truly invisible web	Truly invisible web*
	Pseudo-invisible web

Classification

- **Opaque web** disappears from the viewpoint of this kind of crawlers, excepting for those that have size limits
- **Private web** is different from this perspective because of program options, that sometimes let the user to introduce a password if it is known, or to decide if they want to respect the honor robot protocol or not
- **Pseudo-invisible web** consists of files that looks invisible to the user, but not for crawlers, because they are not correctly displayed or showed in the final reports generated by the software

Truly invisible web

- Disconnected pages
- Non-textual data: images, audio, video
- PDF, PS, Flash, Shockwave, executable and compressed files
- Databases
  - User interaction absent
- Dynamic web pages

Conclusions

- There is a different invisible web for commercial and academic crawlers, but a parallel classification can be drawn
- There are great differences even between crawlers, specially on dynamic web pages
- Technical problems are very similar for search engines and crawlers
- Commercial and academic crawlers must beat some technical limitations
- **Invisible web** can be defined, from a crawler perspective, as the part of the web that commercial and academic crawlers can not see