University of Tunis II
Ecole Nationale des Sciences de l'Informatique

Hannibal - A Search Engine

With a plug-in in Arabic

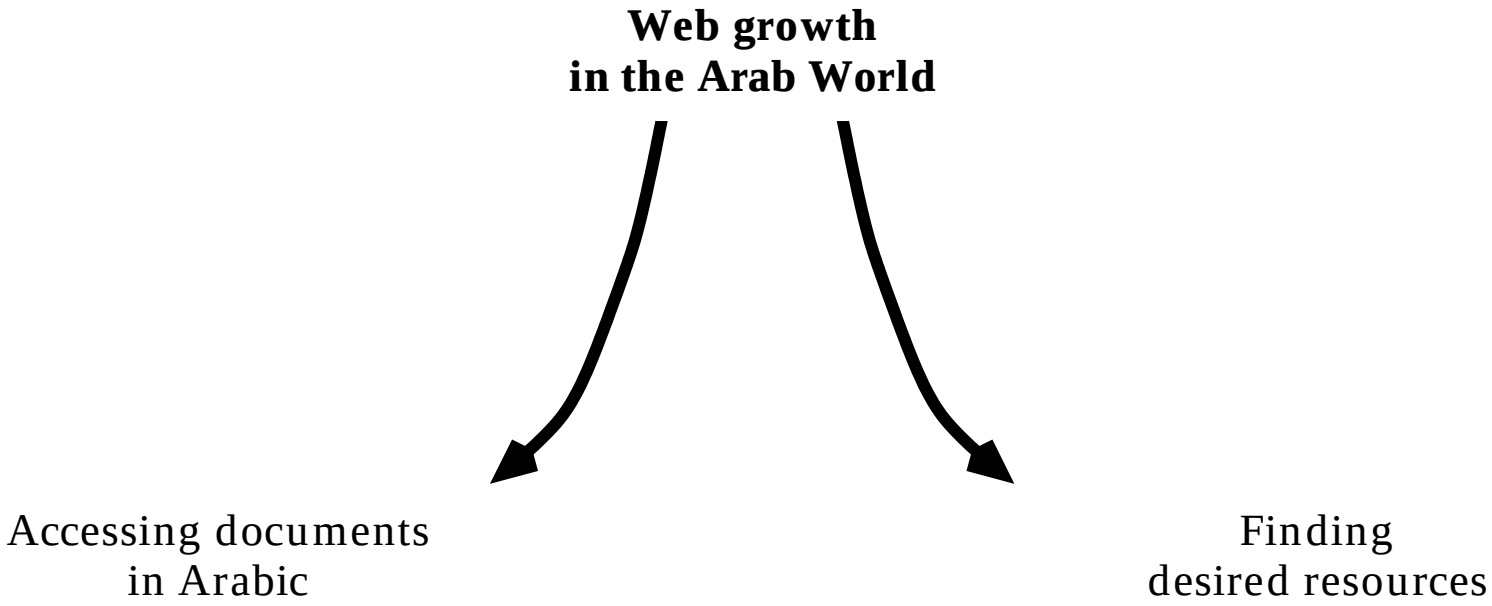
by Imed CHIHI

Supervised by

Mohamed Said OUERGHI, ENSI

Moez SOUABNI, MRS

The problem


Web growth
in the Arab World
Accessing documents
in Arabic
Finding
desired resources

Plan

Part 1: Problems of the Arabic language,
A more general issue: internationalization,
Possible approaches,
Technology used.

Part 2: A search service,
Functions.

Conclusion & perspectives

Part 1: Arabic language support

Problem categories

1. character representation,
2. inconsistencies due to multilingual typesetting,
3. satisfy the needs without breaking with the standards.

Problems of the Arabic language

Multiple character sets,

Typesetting methods not well established,

Complex morphology,

Complex physiology,

Bidirectionnality.

Internationalization issues

US-ASCII used to be ***the*** character set

New tendencies:

- ISO / IEC 10646, then Unicode,
- Enhancements to HTML,
- Using MIME with HTTP,
- ISO 10646 as the previleged character set,
- RFC 1556 for bi-directionnality.

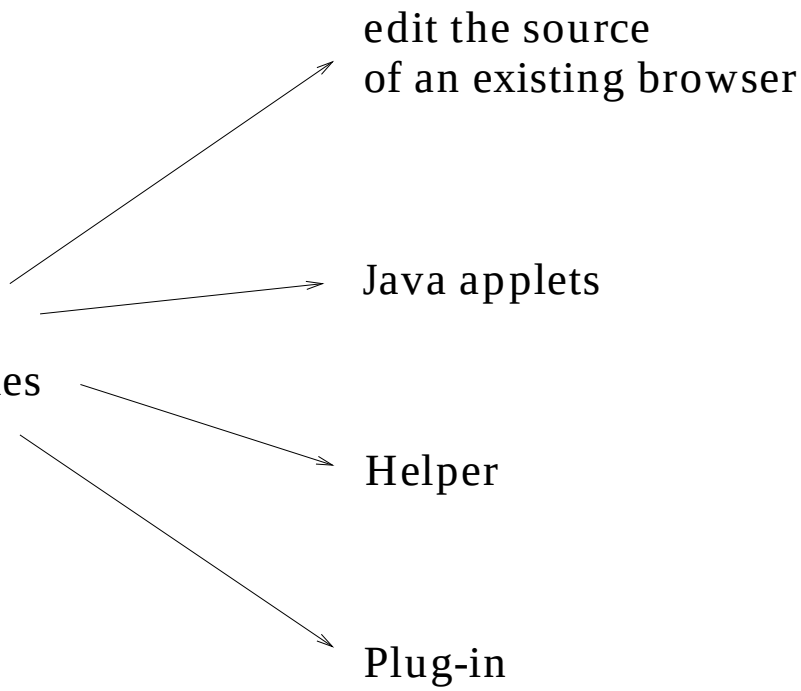
→ Needs ***intelligent*** user agents.

User agents

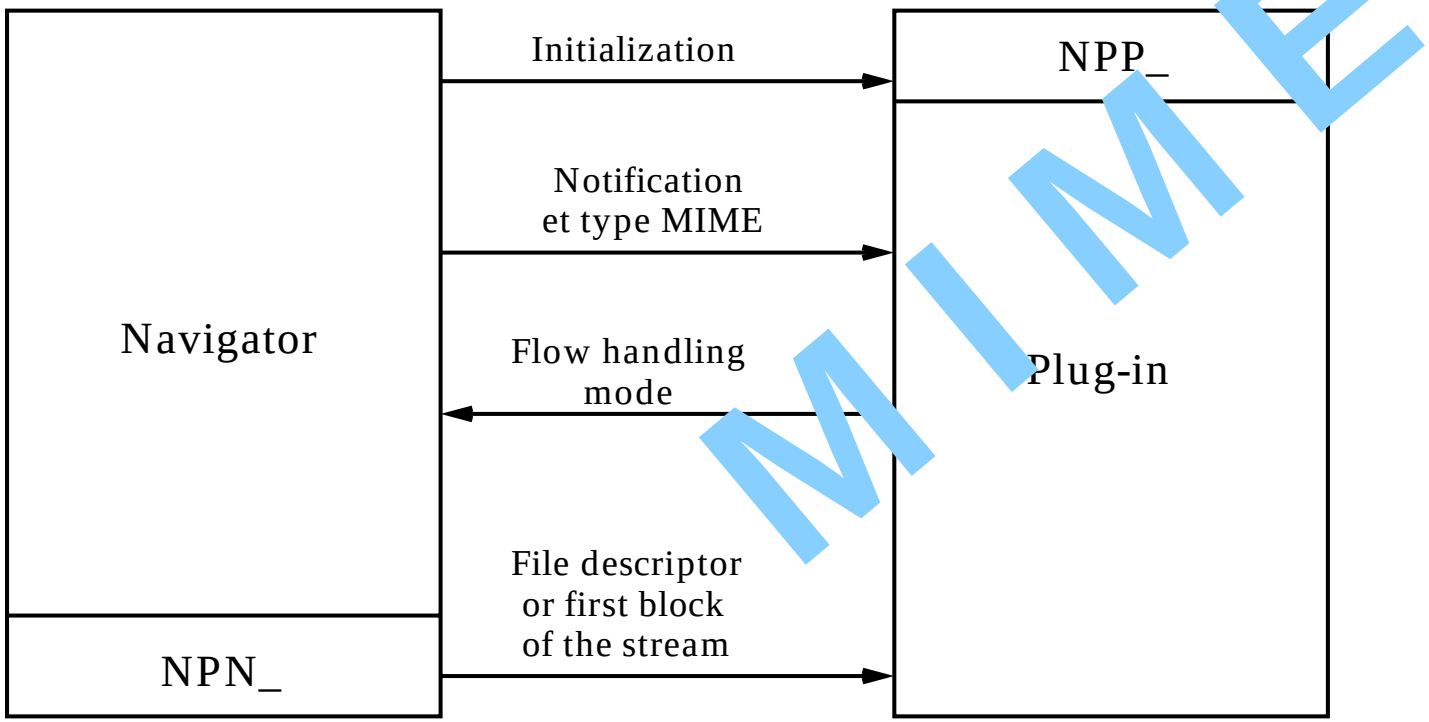
	CP 1256	ISIRI 3342	ISO 8895-6	UTF	Unix	Mac	Windows
Sakhr Sindbad			X	X			X
Alis Tango	X	X	X	X			X
Accent Multilingual Mosaic	X		X				X
Netscape Communicator				X	X	X	X
LangBox AraMosaic	X	X	X		X		
PMosaic		X			X		
Internet Explorer				X	X	X	X

Solving the problem

- "Rely" on Unicode
- Adding functionnalities



Plug-ins technology



Registered MIME type

text / html; charset=iso-8859-6

- a too large plug-in,
- not the objective,
- the solution does exist

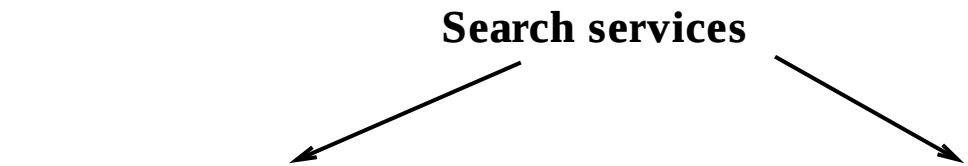
→ text / plain: unformatted text

The charset:

CP 1256	proprietary
ASMO_449 et ISIRI 3342	no Latin glyphs

ISO 8859-6 → ISO 8859-6-i
ISO 8859-6 → ISO 8859-6-e
ISO 8859-6 → ISO 8859-6

Part 2: The search engine



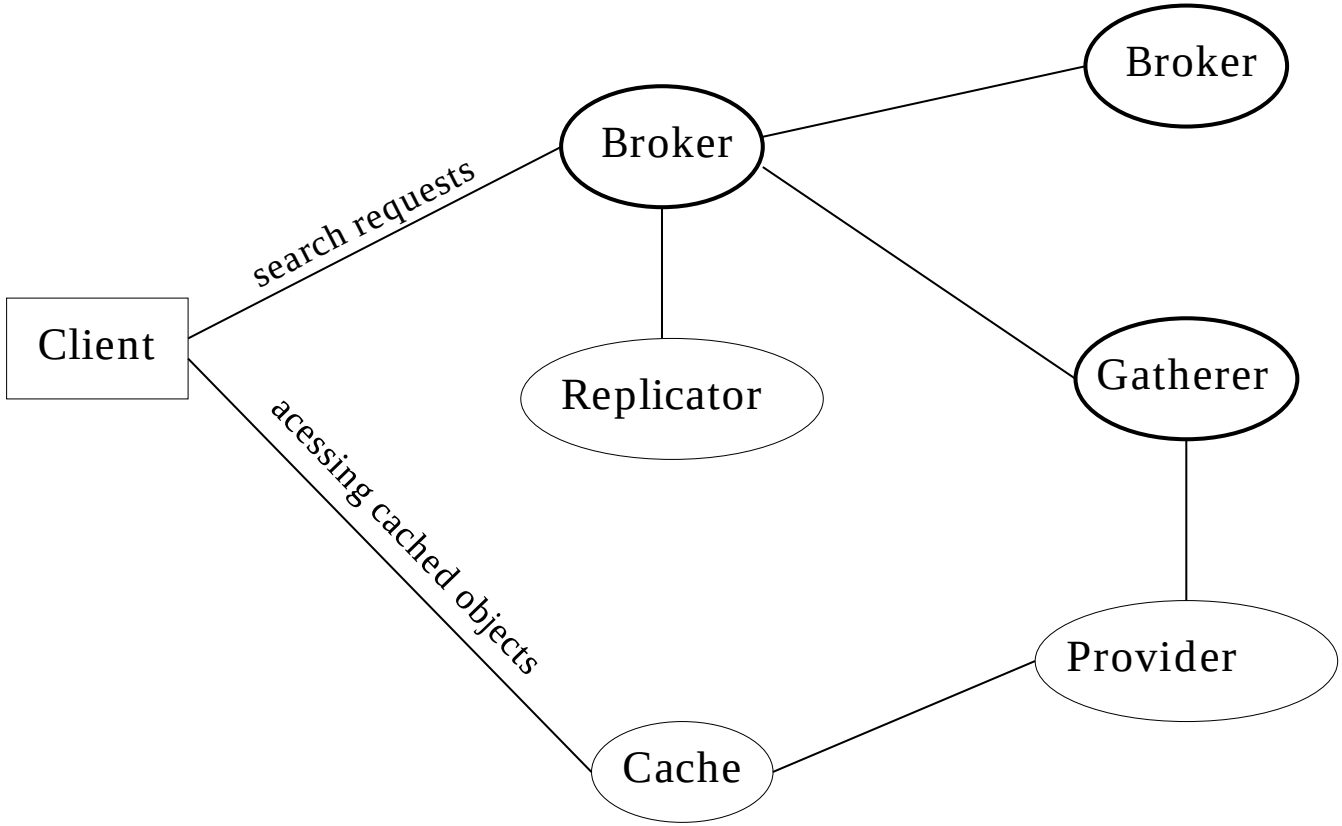
Search engines

- automatic indexing,
- automatic updating,
- average quality index.

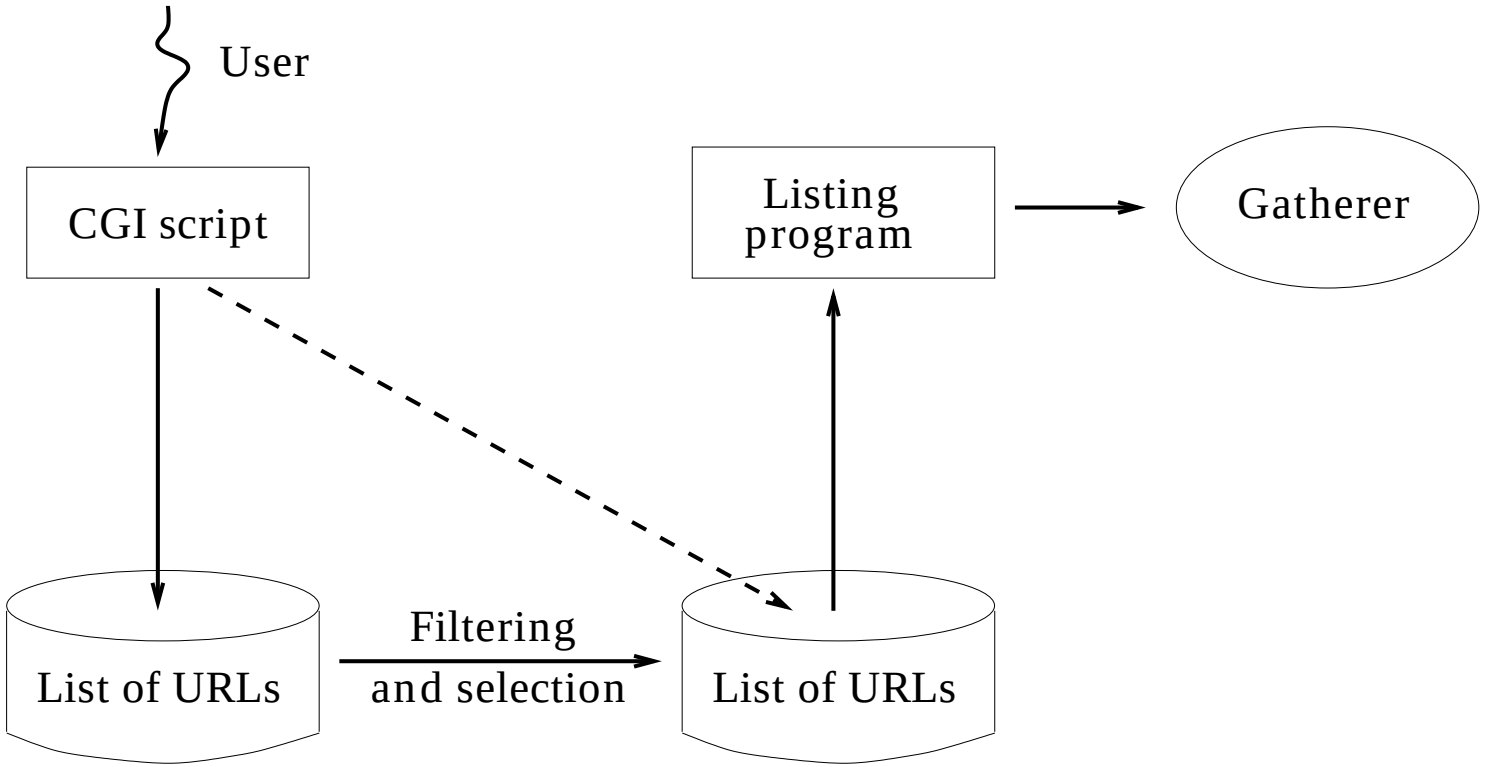
Directories

- manual indexing,
- absolescences are likely,
- high quality index.

The Harvest system



The *Add URL* service



The *Advanced Query* service

Case-sensitivity toggelling,

Search on word boundaries,

Searches with spelling errors,

Display of matched lines.

Conclusion

- What I did:
  - An overview of the existing standards and technologies,
  - A tool for Arabic documents visualization,
  - A functionnal search service.
- The ultimate solution for Arabic:

  Unicode

Perspectives

- Plug-in
  - ports to other platforms,
  - adding other encodings support (especially UTF-8),
  - promote author systems in Arabic.

- The search service
  - distribute the gathering and brokering processus,
  - index documents in Arabic.