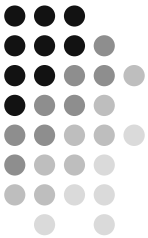
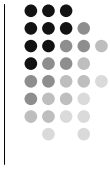
Hyperlinks analysis in Multilingual Web Applications

Filippo Ricca
ITC-irst
Centro per la Ricerca Scientifica e Tecnologica
ricca@itc.it

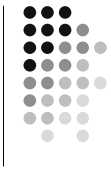


Multilingual Web applications



- To be accessible to a larger audience, Web applications are often required to be multilingual.
- A multilingual Web application is an application where the information contained in the Web pages is supplied in more than one language (e.g., Italian, English, Chinese, etc.).
- An example of it is the Gateway to the European Union (http://europa.eu.int/) where information is supplied in 20 different languages.

Quality of Web applications



- Quality of Web applications is very important.
- In general the quality delivered is often poor and this is particularly true for legacy multilingual Web applications.
- In multilingual Web sites information and structure should be consistent across languages.

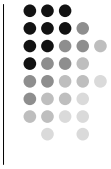
```
it/order.htm
it/payment.htm
en/order.htm
en/payment.htm
```

```
order-it.htm
payment-it.htm
order-it.htm
payment-en.htm
```

```
ordine.htm
pagamento.htm
order.htm
payment.htm
```

In absence of CMS, development of multilingual sites is based on personal (non-standard) choices, and alignment during maintenance is achieved manually. This is particularly true for legacy Web applications.

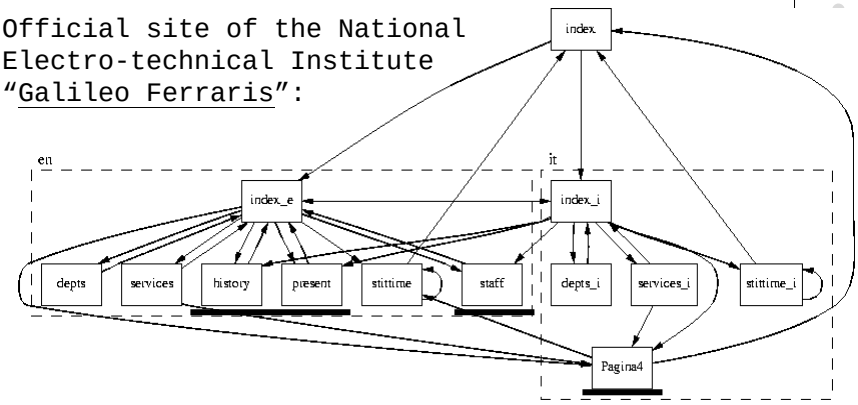
Multilingual Web inconsistencies



Absence of pages in some languages (missing translations)
Differences in the page structure in different languages
Missing information
Parts of the page not translated
Different internal or external hyperlinks.
Incorrect cross-language hyperlinks.
Incomplete change propagation across languages: pages not up-to-date.

Missing translation

Official site of the National Electro-technical Institute "Galileo Ferraris":

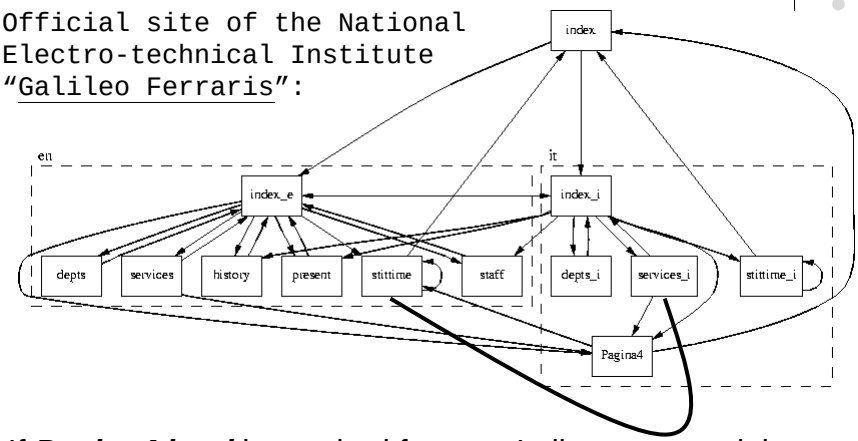


Pagina4.html is available only in Italian.

History.html, present.html and ***staff.html*** are only in English.

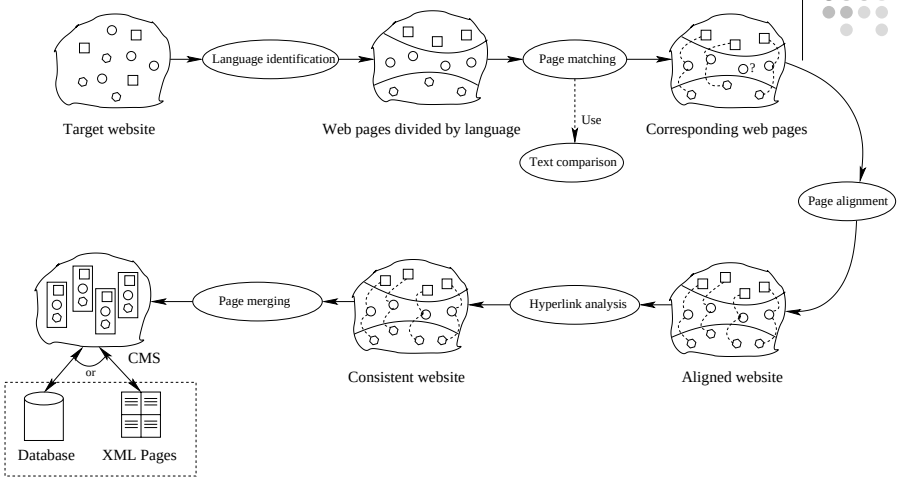
Unintend language switch

Official site of the National Electro-technical Institute "Galileo Ferraris":



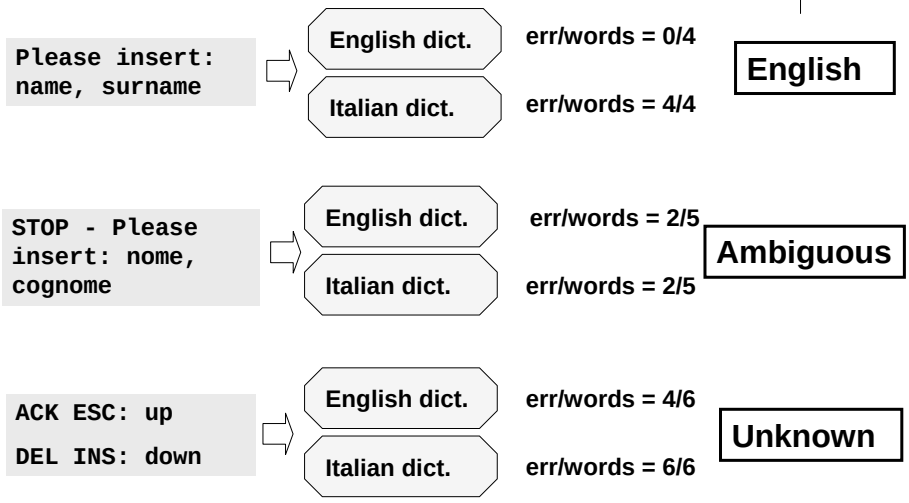
If ***Pagina4.html*** is reached from an Italian page and the standard time is accessed the English version of the page is showed.

Alignment toolkit



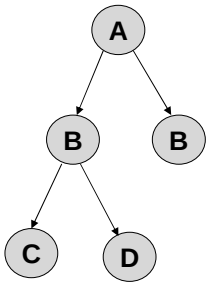
- To cope with Web inconsistencies we devised a number of algorithms implemented and tested in a toolkit.

Language identification

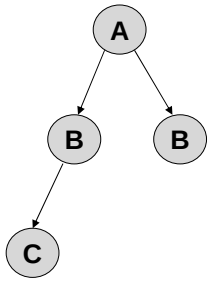


Page matching: structure

```
<A>
<B>
…text1…
<C>…text2…</C>
<D>…text3…</D>
</B>
<B>
…text4…
</B>
</A>
```




```
<A>
<B>
…text5…
<C>…text6…</C>
</B>
<B>
…text7…
</B>
</A>
```



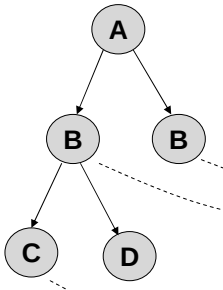
(A, B, C, /C, D, /D, /B, B, /B, /A) **(A, B, C, /C, /B, B, /B, /A)**

del(D), del(/D)

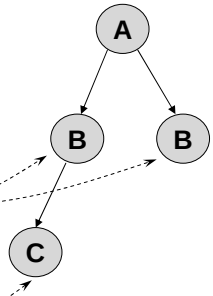
dist = 2

Edit distance: content

```
<A>
<B>
…text1…
<C>…text2…</C>
<D>…text3…</D>
</B>
<B>
…text4…
</B>
</A>
```




```
<A>
<B>
…text5…
<C>…text6…</C>
</B>
<B>
…text7…
</B>
</A>
```



transl(text2, text6) = false

(A, B, C, /C, D, /D, /B, B, /B, /A) **(A, B, C, /C, /B, B, /B, /A)**

del(D), del(/D), del(text2), ins(transl(text6))

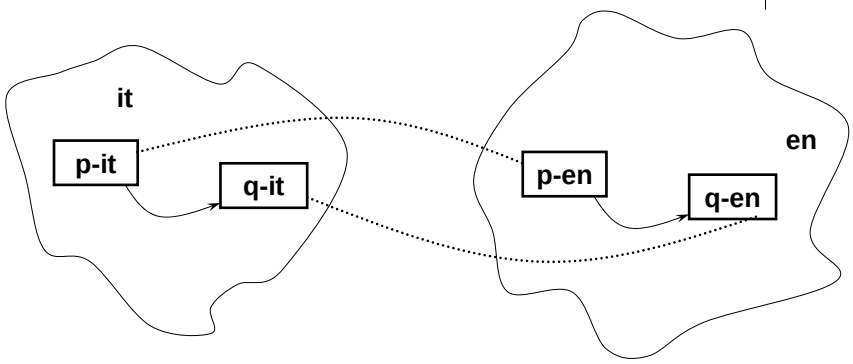
dist = 4

Page alignment

- Output of page matching: pairs of corresponding pages in different languages.
- *Edit script*: edit operations to be performed to align corresponding pages (including text translations).

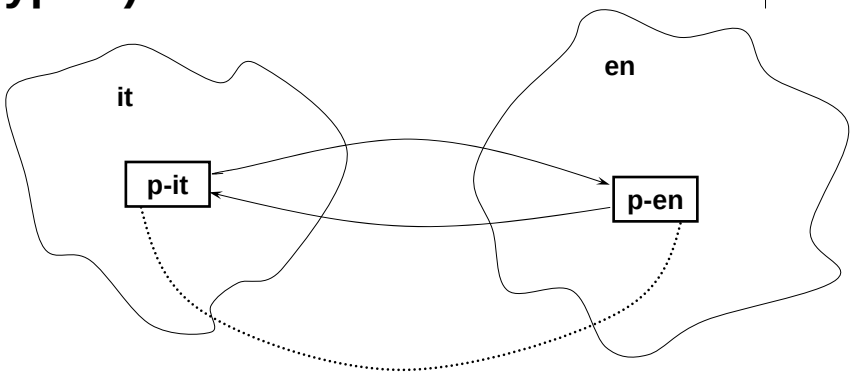
The Web site developer can update pages making them cross-language consistent

Internal hyperlink inconsistency



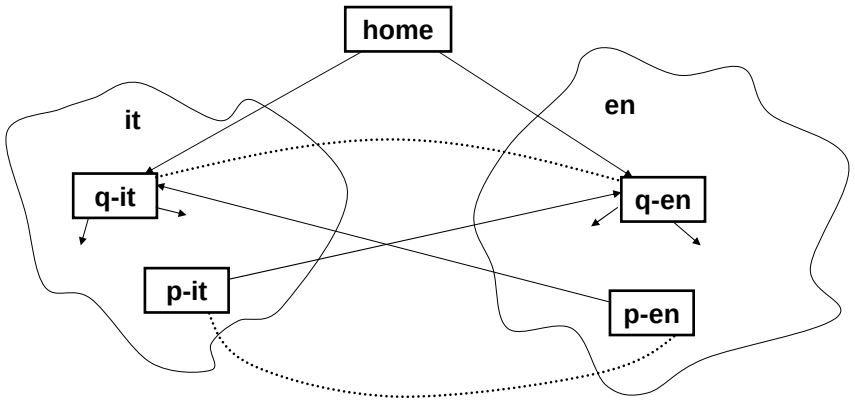
Corresponding pages **p-en**, **q-en** should also be connected by a hyperlink.

Cross-language hyperlinks: external hyperlink inconsistencies (type 1)



A hyperlink from **`p-en`** to **`p-it`** should also exist.

Cross-language hyperlinks: external hyperlink inconsistencies (type 2)

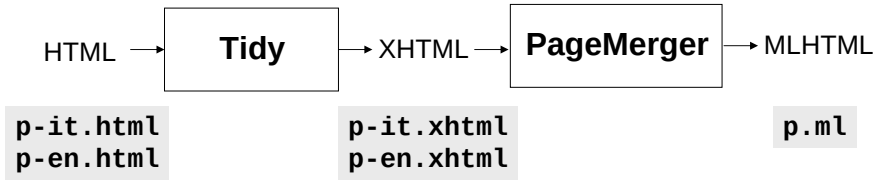


A hyperlink from **`p-en`** to **`q-it`** should also exist.

```
FOR EACH a_lang1 → b_lang1 ∈ PORTION1 {
   a_lang2 = pageMatching(a_lang1);
   b_lang2 = pageMatching (b_lang1);
   IF a_lang2 → b_lang2 ∉ PORTION2
     Int_Set = Int_Set ∪ {a_lang2 → b_lang2}
}
FOR EACH a_lang2 → b_lang2 ∈ PORTION2 {
   a_lang1 = pageMatching(a_lang2);
   b_lang1 = pageMatching (b_lang2);
   IF a_lang1 → b_lang1 ∉ PORTION1
     Int_Set = Int_Set ∪ {a_lang1 → b_lang1};
}
FOR EACH a_lang1 → b_lang2 ∈ ALL {
 a_lang2 = pageMatching(a_lang1);
 b_lang1  = pageMatching (b_lang2);
 IF b_lang2 == a_lang2 { // Type 1
  IF b_lang2 → a_lang1 ∉ ALL
   Ext1_Set = Ext1_Set ∪ {b_lang2 → a_lang1};
 } ELSE { // Type 2
  IF a_lang2 → b_lang1 ∉ ALL
   Ext2_Set = Ext2_Set ∪ {a_lang2 → b_lang1};
}
```

Pseudo-code

Page merging



where **p-it.html** and **p-en.html** are aligned.

MLHTML: Multilingual XHTML

- MLHTML is an XML representation of multilingual Web pages
- Information in all languages is inserted into a single file
- Document format and structure are shared among languages
- The multilingual content is inserted inside adjacent lines

```
<title>
<ml lang="en"> Publication List </ml>
<ml lang="it"> Lista delle pubblicazioni </ml>
</title>
```

Experimental Results

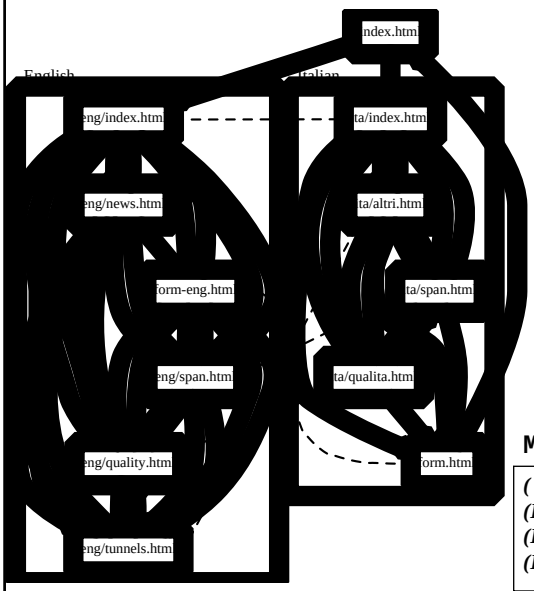
Web sites	Pages	LOC	Links
www.sitaserre.com	18	2732	252
www.alberghiero-longarone.it	25	8028	182
www.mediaelettra.com	37	29954	2144
www.artifer.com	54	10827	1003
www.toscana-toskana.de	68	13734	1240
www.trevirtu.com	81	17522	870
www.gigirosso.com	84	11195	1689
www.agenziarossi.com	125	55587	558
www.viniesaporidipuglia.com	210	28036	2418
www.olio-oliva.it	275	25425	809
Total	977	203040	11165

- 10 Web applications chosen among Italy-based firms and educational institutions has been downloaded by means of the **ReWeb** tool.

Hyperlinks inconsistencies reveled

Web sites	Int	Ext 1	Ext 2
www.sitaserre.com	4/36	0/0	0/0
www.alberghiero-longarone.it	10/36	0/8	7/21
www.mediaelettra.com	23/648	1/3	1/27
www.artifer.com	0/72	0/0	0/0
www.toscana-toskana.de	9/90	2/8	6/28
www.trevirtu.com	9/377	1/51	9/9
www.gigirosso.com	0/342	0/0	0/0
www.agenziarossi.com	16/301	0/0	0/1
www.viniesaporidipuglia.com	24/2192	0/2	7/21
www.olio-oliva.it	18/661	0/2	12/13
Total	113/4755	4/74	42/120

www.sitaserre.com

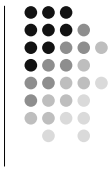


Four outgoing links of form-eng.html have no corresponding links outgoing from form.html.

Missing hyperlinks

(Int, form.html ----> ita/index.html, Italian)
(Int, form.html ----> ita/qualita.html, Italian)
(Int, form.html ----> ita/span.html, Italian)
(Int, form.html ----> ita/altri.html, Italian)

Conclusions and future work



- We have explained in detail a module of the Alignment Toolkit: the **Hyperlinks analyzer.**
- The Alignment Toolkit helps the Web maintainer to re-structure an existing Web site to ensure consistency among its multilingual portions. The result of this process is a Content Management System, which internally stores multilingual content and page structure.
- Before creating the Content Management System the Hyperlinks analyzer is applied. The result of this module is a list of 'missing links' able to make the hyperlinks structure of the target Web application consistent.
- A preliminary evaluation of the Hyperlinks analyzer has been conducted on a sample of 10 multilingual Web sites: in 8 cases out of 10 at least an internal inconsistency or an external one is present.
- As a future work we plan to conduct further experiments on dynamic Web applications.