

# BIB<sub>T</sub>E<sub>X</sub> Bibliography Index Maker (Bibim)

Primavera 2010

# Índex

## Objectius

## Extracció de referències

- PDF to text

- Cerca de referències

- Extracció d'informació

## Generació de regles

- Tipus de Regles

- Avaluació

## Demo

## Conclusions

# Objectius

## Motivació

### Guió típic d'un treball de recerca:

1. Lectura d'articles
2. S'acumulen els fitxers (PDF) en un directori
3. A l'hora d'escriure, fa falta tenir un índex bibliogràfic per poder citar

# Objectius

## Objectiu

Un únic objectiu principal:

- ▶ La creació d'índexs bibliogràfics de forma automàtica

Però per aconseguir-lo necessitarem:

- ▶ Cercar referències a Internet  
(Biblioteques digitals)
- ▶ Extreure la informació de les pàgines HTML
- ▶ Crear regles d'extracció automàticament

# Objectius

## Objectiu

Un únic objectiu principal:

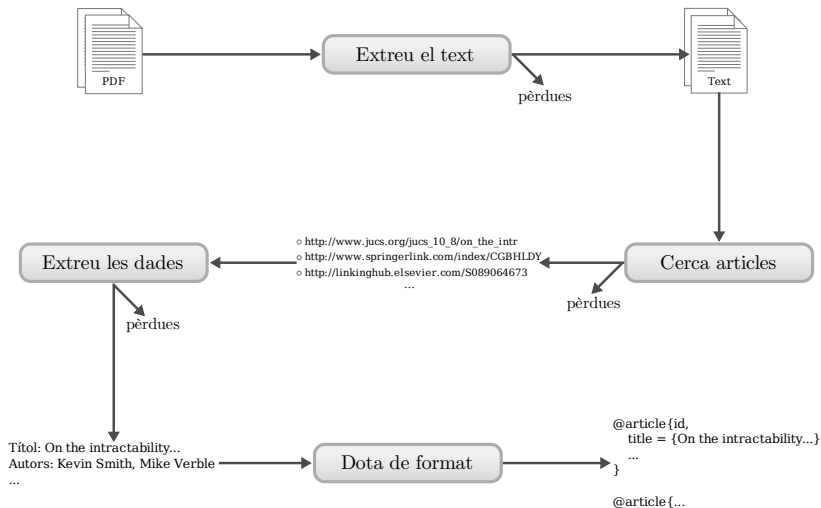
- ▶ La creació d'índexs bibliogràfics de forma automàtica

Però per aconseguir-lo necessitarem:

- ▶ Cercar referències a Internet  
(Biblioteques digitals)
- ▶ Extreure la informació de les pàgines HTML
- ▶ Crear regles d'extracció automàticament

# Extracció de referències

## Esquema d'extracció



# Extracció de referències

*PDF to text*

Quant a l'extracció de text dels fitxers PDFs:

- ▶ S'utilitza l'eina *xPDF*
- ▶ Resultats relativament bons...
- ▶ ...però continua tenint força problemes

# Extracció de referències

## Cerca de referències

### Passos a seguir per cercar referències:

- ▶ Genera consultes a partir del text

`([\w()?!]+\ ){\min,max}`

*“and slurry methods and have been tested in the selective”*

- ▶ Obté els resultats de les consultes (*Google, Bing, Yahoo*)

- ▶ `www.springerlink.com/index/G4588X...`

- ▶ `www.ingentaconnect.com/content/klu/ca...`

- ▶ Filtra i ordena els resultats

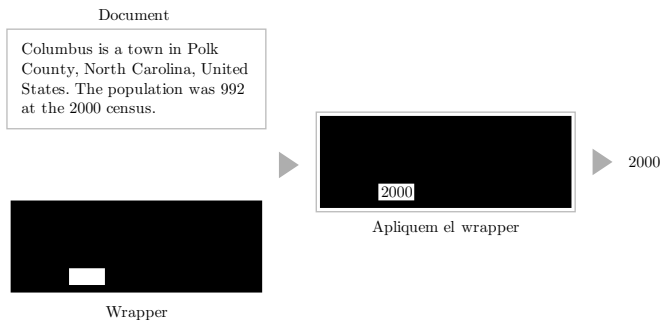


# Extracció de referències

## Extracció d'informació

### Dos tipus de regles d'extracció d'informació:

- ▶ *Reference Wrappers*
- ▶ *Field Wrappers*

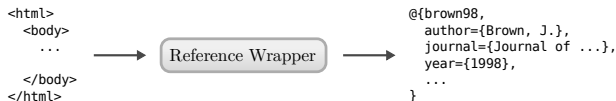


# Extracció de referències

## Reference Wrappers

### Característiques:

- ▶ Extreuen referències senceres (BIB<sub>T</sub>E<sub>X</sub>)
- ▶ Només se'n necessita un per cada biblioteca digital
- ▶ Implementats manualment
- ▶ Els resultats solen ser bons

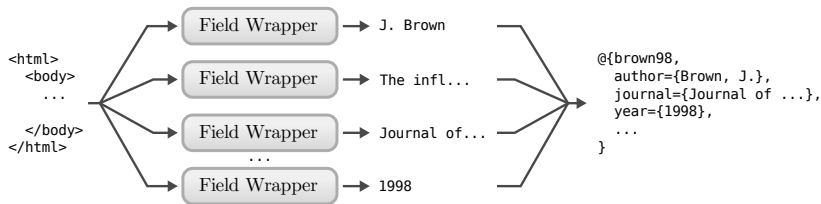


# Extracció de referències

## Field Wrappers

### Característiques:

- ▶ Especialitzats en extreure un sol camp
- ▶ Se'n necessita un per cada camp i per cada biblioteca
- ▶ Implementats automàticament



# Extracció de referències

## Validació

### Per què:

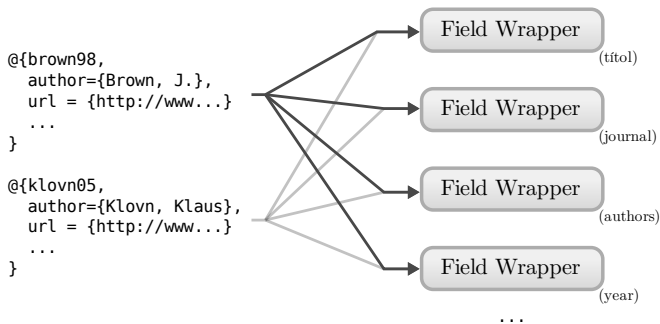
- ▶ Informar a l'usuari
- ▶ Utilitzar les referències per re-generar regles

### Com ho fem:

- ▶ Camps com ara el títol i autors: es comprova que es troben dins del document PDF
- ▶ Camps com ara el número de pàgines o any: es mira que compleixin una expressió regular.
- ▶ Cada camp té un pes sobre la validesa final

# Generació de regles

## Esquema

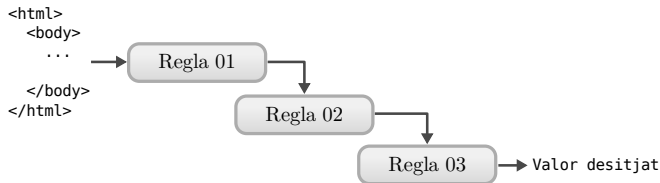


- S'utilitzen (poques) referències com a exemples.
- Corresponen a la mateixa biblioteca digital

## Generació de regles

## Estructura dels *Field Wrappers*

Els *field wrappers* es caracteritzen per:

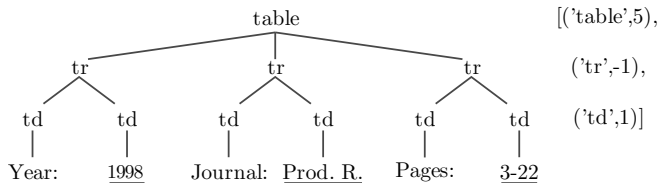


# Generació de regles

## Tipus de regles

### *Path Rule:*

- Localitza elements del document HTML
- El patró consisteix en una ruta dins de l'arbre HTML



# Generació de regles

## Tipus de regles

### *Regex Rule:*

- ▶ Localitza valors dins d'una cadena de caràcters.
- ▶ El patró és una expressió regular per extreure el valor desitjat

Text:        **Vol. 27, pp. 1204–1209.**

Patró:        **(?:.\*) , pp. (.\*) .**

Resultat:    **1204–1209**



# Generació de regles

## Altres regles

### *Separators Regex Rule:*

- ▶ Separa valors continguts en un mateix text
- ▶ El patró és un conjunt de cadenes que actuen de separadors

Text:       **Value A, Value B and Value C**

Patró:       **", " i " and "**

Resultat:   **[Value A, Value B, Value C]**

# Generació de regles

## Altres regles

### *MultiValue Regex Rule:*

- ▶ *Regex rule* que actua sobre diferents valors

### *Person Rule:*

- ▶ Separen noms en els camps: `last_name`, `middle_name` i `first_name`. Sense patró.

Nom:       **Anders And**

Resultat:   [ **"And"**,   **"",** **"Anders"** ]

# Generació de regles

## Altres regles

### *MultiValue Regex Rule:*

- ▶ *Regex rule* que actua sobre diferents valors

### *Person Rule:*

- ▶ Separen noms en els camps: `last_name`, `middle_name` i `first_name`. Sense patró.

Nom:       **Anders And**

Resultat:   [**"And"**, **"", "Anders"**]

# Generació de regles

## Avaluació de *Wrappers*

Quant a l'avaluació de *wrappers*:

- ▶ Els millors *wrappers* s'han de provar primer
- ▶ Per escollir un ordre inicial, s'apliquen sobre els mateixos exemples
- ▶ Cada vegada que s'utilitza un *wrapper* se li dóna un vot
- ▶ Els bons pugen i els dolents baixen

# Demo

# Conclusions

- ▶ Pèrdues a cada pas
- ▶ L'èxit depèn completament de la qualitat dels *wrappers* disponibles
- ▶ Errors fàcilment corregibles
- ▶ La part de generació i extracció, aplicable a qualsevol altre domini
- ▶ Molt per fer i millorar

## Q&A