Servei de recerca automatitzada de continguts

Abraham Sánchez Pino

15 Gener 2012

2on cicle EI

Consultor: Jordi Ferrer Duran

Presentació del projecte

- Aquest projecte es basa en el disseny i realització d'un servei de recerca automatitzada de continguts sindicats en web
- Els formats suportats són RSS (Really Simple Syndication) i Atom
- Filtrat de la informació mitjançant regles de selecció
- Forta motivació tecnològica
- Desenvolupat enterament amb tecnologies de codi obert

Formats de sindicació

- **Sindicar** és posar un contingut a disposició dels usuaris, ja siguin persones o sistemes informàtics
- **RSS**
  - Inventat per Netscape el 1999
  - Versió actual: 2.0 ➜ format de sindicació més estès globalment
  - Extensible mitjançant etiquetes amb espais de noms diferents
- **Atom**
  - Proposat com a estàndard d'internet en el IETF
  - Intenta eliminar les ambigüitats de RSS
  - Extensible mitjançant mòduls. Més guiat i millor dissenyat.

Objectius principals

- **Aprofundir en el coneixement dels formats de sindicació**
- Oferir una experiència basada en tecnologies web punteres i obertes
- Familiaritzar-se amb les tecnologies XML estàndard, com XSL o Xpath i amb conceptes interessants pel món dels compiladors com la metaprogramació o la reflexió
- Posar en pràctica i assimilar els patrons de disseny actuals (MVC, DRY, CoC, RESTful ...)
- Iniciar-se en el món del desenvolupament dirigit per, o orientat a, tests

Planificació i procés de desenvolupament

- Projecte amb molta vocació tecnològica i de recerca
- Procés de documentació prèvia i experimentació alt; més del previst
- **Avantatges:**
  - codi molt optimitzat
  - simplificació de la implementació
- Metodologia orientada a test (Test Driven Development).
  - **Qué implica?**
    - Aplicació més mantenible i robusta, millor testejada
    - Dissenys més simples i confiables

Components

- L’aplicació final presenta els següents components o mòduls principals:
  - **Lectura de fonts**
  - **Exportació de fonts**
  - **Regles de selecció (de fonts)**
  - **Interfície web**
  - **Base de dades orientada a documents**

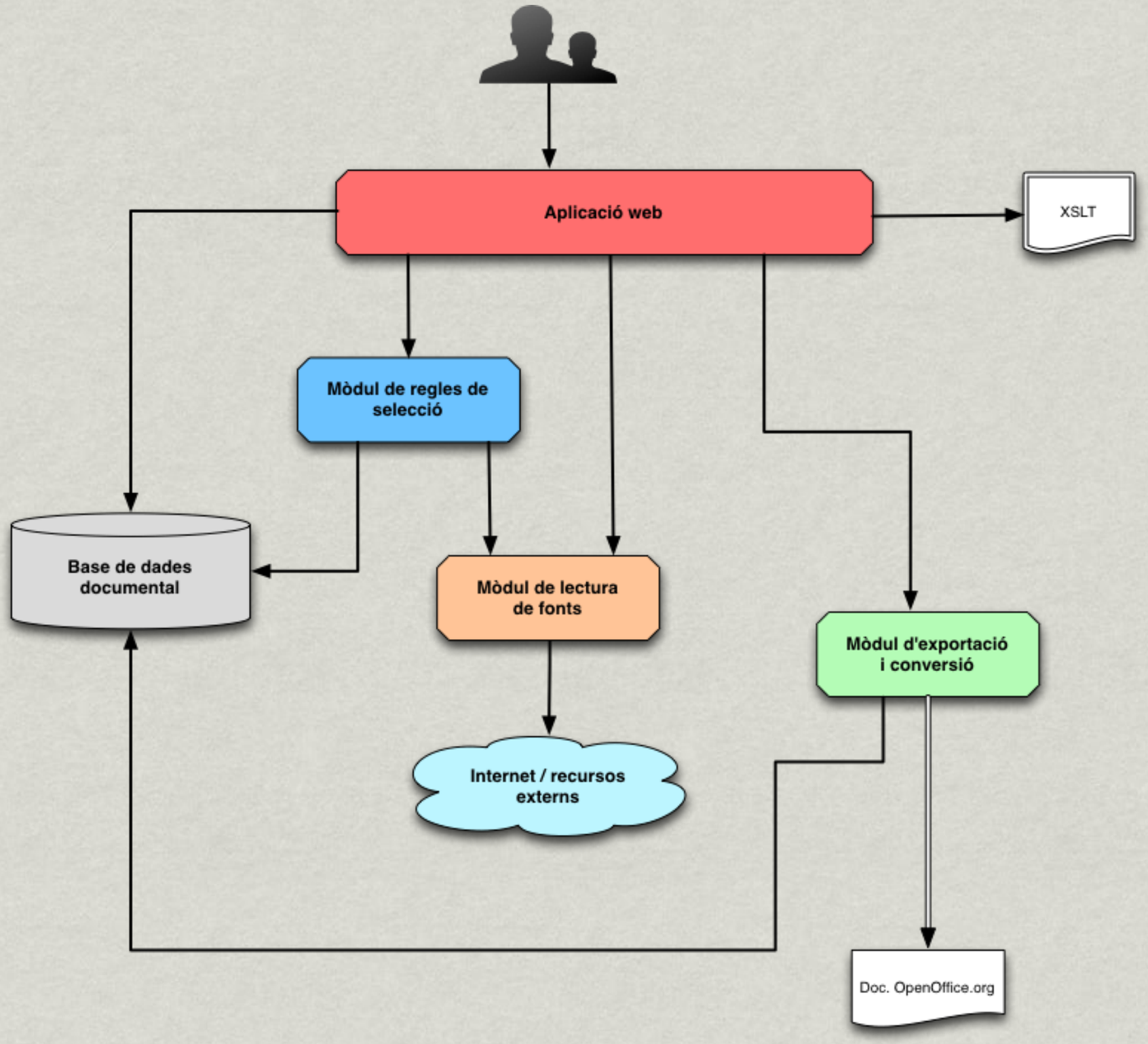
Característiques generals

- Lectura de fonts RSS i Atom ➜ Conversió a format intern (propi)
- Emmagatzemament i indexació dels articles
- Interfície web per accedir a les funcionalitats principals de les fonts
- Exportació dels articles a OpenDocument (.odt)
- Procés desatès (cron) que actualitza automàticament les fonts

Regles de selecció

- Cada font està associada a una regla de selecció que aporta filtres pels quals limitar els articles
- **Temporalitat**: data de publicació
- **Contingut**: paraules incloses en el contingut
- **Idioma**: filtre per idioma
- **Autoria**: filtre per un autor concret
- **Categories**: l'article pertany a una o vàries categories concretes

Arquitectura general



Principals decisions

- Cal establir un subconjunt bàsic d'atributs comuns entre RSS 2.0 i Atom
- Les fonts es converteixen a un format homogeni i s'emmagatzemem en una base de dades no relacional
- Visualització d'articles mitjançant XSLT
- Cada funcionalitat està associada a un joc de proves

Tecnologia

- Llenguatge de programació **Ruby** ➜ Dinàmic, expressiu, orientat a objectes i extensible
- API de tractament XML: **Nokogiri**
- Base de dades NoSQL: **MongoDB**
- *Framework* web: **Ruby on Rails** ➜ Ús de patrons i tècniques de programació avantguardistes
- *Testing* amb **RSpec** i **Capybara**

Patrons de disseny

- MVC (**M**odel - **V**ista - **C**ontrolador)
- CoC (**C**onvention **o**ver **C**onfiguration)
- DRY (**D**on’t **R**epeat **Y**ourself)
- ActiveRecord
- IoC (**I**nversion **o**f **C**ontrol)

Principals reptes

- Gran quantitat de metodologies, tècniques i patrons que cal dominar
- Les implantacions de RSS 2.0 són molt diverses ➜ Exemples
- Assegurar la compatibilitat de qualsevol tipus de font és molt complicat
- Atom està, malauradament, poc estès en comparació
- La integració entre solucions tecnològiques a vegades ha presentat problemes inesperats en el desenvolupament que l'han endarrerit inexorablement.

Resultats obtinguts

- L'aplicació, si bé li manquen moltes coses per ser considerada en un entorn corporatiu o professional, ofereix la funcionalitat bàsica desitjada
- Tots els mòduls funcionals plantejats han estat implementats, tot i que en algun d'ells s'ha hagut de descartar algunes funcionalitats menors

Conclusions

- Els formats de sindicació en l'actualitat estan dominats per RSS 2.0
- Existeix un nivell força elevat de disgregació en l'ús del format. Les extensions no estan regulades ni estandarditzades
- Atom podria ser una solució al problema però no té (encara?) la difusió necessària
- L'**estudi de les tecnologies** aplicades ha estat crucial pel bon desenvolupament del projecte. Plantejar-se no fer-ne ús multiplicaria el temps d'implementació
- Hem vist com es desenvolupa una aplicació web amb un *framework* modern i actual
- Els **patrons de disseny** realment aporten un valor afegit al projecte. És quan s'en fa ús d'ells quan ens podem adonar de la seva utilitat.

Aportacions i millores futures

- Suportar i discriminar entre diferents tipus de *feeds* (*enclosures*, *podcasts*, *photocasts*, etc.)
- Augmentar compatibilitat amb les extensions dels formats
- Oferir més opcions d'exportació ➜ PDF, MS Word, etc.
- Potenciar aplicació web ➜ Registre d'usuaris, més possibilitats

Gràcies!

- Abraham Sánchez ➜ asanchezpi@uoc.edu