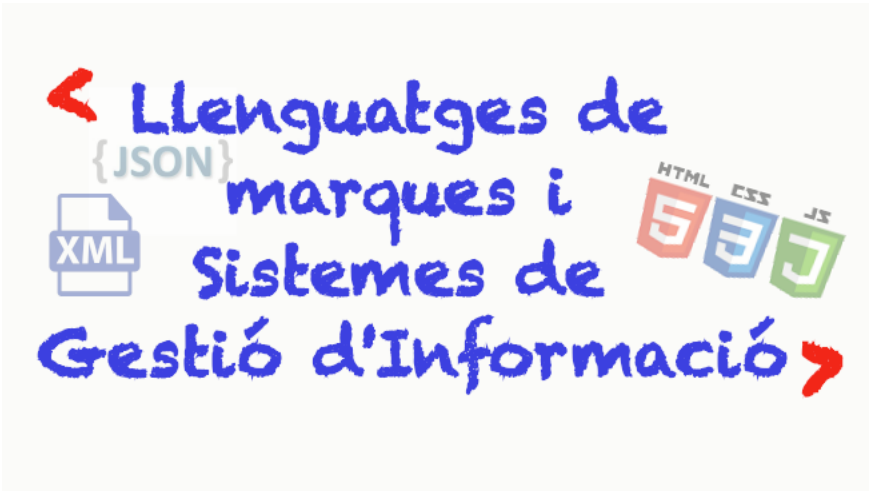
Llenguatges de marques per a la sindicació de continguts

IES Sant Vicent Ferrer
Algemesí



Contenido

1. Introducció a la sindicació de continguts

La primera manera de sindicació que va haver va ser la integració de notícies d'una pàgina dins d'una altra per mitjà d'algun tipus de programa que obtenia la informació cercant per dins del contingut HTML. Però aquests programes tenien el problema que sovint quedaven obsolets, ja que qualsevol canvi en la pàgina original podia fer que deixaren de funcionar correctament.

A pesar del que puga semblar per l'abundància de "decoració" que hi ha en les pàgines web, el més important és el contingut. La informació continguda en els articles, els fitxers, etc., és el que fa que els visitants tornen o no.

Amb el sistema tradicional de navegació per Internet, per a un usuari era molt important obtenir els enllaços dels llocs web que li interessaven i emmagatzemar-los d'alguna manera per poder a ells tornar ràpidament. Si es volien seguir els canvis en les pàgines web l'única manera que hi havia era anar visitant de tant en tant la pàgina per comprovar si hi havia novetats.

L'aparició del que es va conèixer com a Web 2.0 va complicar les coses. El Web es va emplenar d'una gran quantitat de blogs i pàgines que publicaven informació, i visitar-les totes per veure si hi havia canvis va passar a requerir molt de temps. **S'havia d'optimitzar d'alguna manera aquesta tasca.**

L'aparició dels **sistemes estàndard de sindicació** va fer possible obtenir la informació de les actualitzacions d'un lloc web d'una manera estable per mitjà d'un adreça. La sindicació de continguts va canviar la manera com es recupera la informació. Ja no calia anar a buscar la informació: era la informació la que acudia a l'usuari.

Fent servir la sindicació ja no cal que l'usuari visite les pàgines que li interessen per a veure si hi ha canvis perquè si n'hi ha ja els rebrà. Això comporta un estalvi de temps, ja que no caldrà visitar pàgines per descobrir que no hi ha canvis. Amb la sindicació el disseny de la pàgina original no afecta els programes que busquen informació, ja que la sindicació de continguts està basada en XML i no prioritza la informació d'estil sinó el contingut.

Aquesta està pensada perquè puguen interactuar tant els humans com els programes, i això fa que es puguen dissenyar fàcilment aplicacions que obtinguen la informació de manera automàtica sense que calga cap tipus d'intervenció humana.

La participació de l'humà es limitarà a dir al programa quins llocs ha de vigilar. A més, un cop l'usuari rep la informació en pot fer el que vulga: filtrar-ne els continguts, classificar-la per temes... Per tant, tindrà el control de quina és la informació que vol veure i quina no.

Un altre dels avantatges que aporta la sindicació és inherent a XML. A diferencia del que passa amb HTML, és fàcil interpretar el contingut de la informació que es rep i, per tant, també serà fàcil poder reutilitzar-ne el contingut per fer-hi altres tasques.

A pesar de que la sindicació es veu sovint com un sistema enfocat a detectar novetats en el Web, també s'està fent servir per a mantenir actualitzacions en altres camps. Per exemple, alguns programes d'ordinador fan servir RSS per saber si hi ha actualitzacions noves i d'aquesta manera mantenen els programes actualitzats.

2. "Feeds" o canals

Un canal és un arxiu que conté una versió específica de la informació que s'ha publicat en un lloc web.

En aquest arxiu es troba tota la informació sobre el lloc web i enllaços als seus continguts. El gran avantatge és que en estar basat en XML es pot aconseguir transmetre la informació de manera automatitzada i els receptors la podran interpretar fàcilment.

Per poder obtenir la informació del canal normalment caldrà localitzar el fitxer. Generalment aquests canals estan associats a una pàgina web i accessible per mitjà d'un enllaç. Els enllaços solen estar clarament especificats amb el text RSS, XML o bé fent servir el grup d'icones següent:

Els fitxers dels canals normalment es passaran a programes que seran els que s'encarregaran de recollir periòdicament les actualitzacions de la informació del canal. En la terminologia de sindicació això se sol anomenar subscripció.

L'ús de canals aportarà diferents avantatges als usuaris:

- Com que funciona per subscripció, aquests només rebran les noticies d'interès seu.
- El programa tendirà a donar informació més acurada a les seues preferències i gustos del que ho fan els cercadors generalistes com Google, ja que els canals contenen el resultat d'una cerca.
- La informació es pot classificar i ordenar segons els gustos de l'usuari i consumir-se segons els criteris de preferència.
- En qualsevol moment es pot deixar de seguir un canal sense haver de demanar cap tipus de permís.

Al llarg dels anys s'han desenvolupat diverses tecnologies per crear canals com CDF (channel definition format, desenvolupat per Microsoft), PointCast o Apple MCF (meta content framework), però els llenguatges de creació de canals que s'han fet més populars i que s'han convertit en la manera estàndard de sindicació han estat sobretot **RSS** i **Atom**.

3. RSS

RSS són les sigles que es fan servir per anomenar diferents estàndards molt populars per a la sindicació de continguts que s'han convertit en una manera estàndard d'intercanviar informació al Web.

a. Llenguatge RSS 2.0

Entre tots els canals disponibles, RSS 2.0 és el més usat amb molta diferència, i per tant, la gran majoria dels programes lectors d'RSS el suporten. Una de les característiques que defineixen RSS és que fa honor al seu nom oficial, really simple syndication (sindicació realment senzilla), i és un sistema senzill.

Es tracta d'un sistema que no té cap estructura complexa, en què les etiquetes descriuen el contingut que hi ha en l'element, en què pràcticament no es fan servir els atributs per a res i els espais de noms només es fan servir en les extensions, si n'hi ha.

L'arrel

RSS és un llenguatge XML, de manera que n'ha de complir les normes i, per tant, només té un sol element arrel, <rss>. La funció d'aquest element és simplement informar a qui llegeixi el document que el que està llegint és un canal RSS.

L'element arrel té un dels pocs atributs obligatoris de l'especificació, version, que és necessari per indicar la versió que s'està fent servir. Aquest atribut serveix perquè els programes sàpiguen quina versió d'RSS es fa servir en el document.

```
<rss version="2.0">
       ...
</rss>
```

L'element <channel>

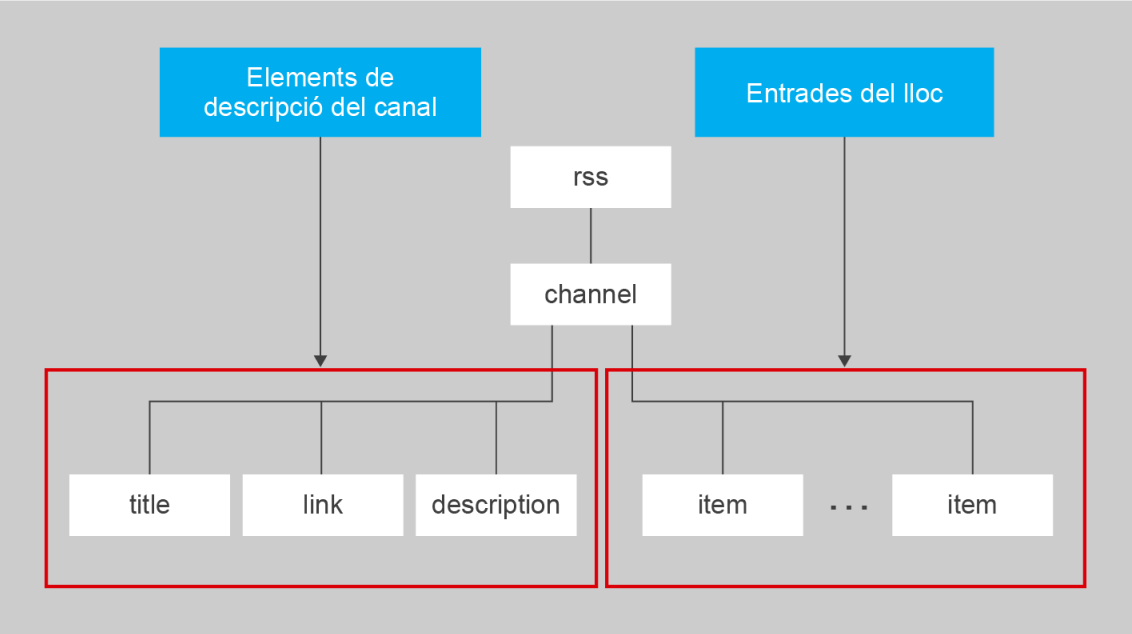
L'arrel només serveix per indicar que el document és de tipus RSS, i el contingut del canal estarà dins de l'únic fill de <rss>, que s'anomena <channel>. L'element <channel> serà el que contindrà totes les etiquetes que aporten informació sobre el canal, i sobretot les que tindran les novetats del lloc. Només pot haver una sola etiqueta <channel> en tot el document RSS.

```
<rss version="2.0">
      <channel>
      ... contingut ...
      </channel>
</rss>
```

Es pot dividir el contingut d'un canal RSS en dos grans grups (figura 1):

1. Un grup d'etiquetes destinades a descriure el canal.

2. Els elements <item>, que són els que contindran el contingut del canal.



Per fer-ho tot més senzill algunes de les etiquetes es repeteixen en els dos grups. Per exemple les etiquetes més importants de <channel>, que són <title>, <link> i <description>, són també les més importants dels elements <item>. Etiquetes per a descriure el canal.

Els primers elements que es troben dins de l'element <channel> estan destinats a donar informació sobre el canal RSS. Aquestes etiquetes no són de contingut ni serà habitual que es produeixen canvis en els seus valors.

Els elements més importants d'aquesta part són els elements <title>, <link> i <description> que són obligatoris en tots els canals RSS.

Aquests són els únics elements obligatoris i, per tant, el document següent seria un document RSS vàlid.

```
<rss version="2.0">
	<channel>
		<title>Canal de llenguatges de marques</title>
		<link>http://ioc.xtec.cat/rss/Marques.html</link>
		<description>Canal per entrar les notes del mòdul 4 d'ASIX</description>
	</channel>
</rss>
```

A part dels elements obligatoris en RSS també n'hi ha uns quants que són voluntaris i que serveixen per donar informació extra sobre el canal:

- language: Especifica l'idioma en que està escrit el canal.
- copyright: Informació sobre el copyright del contingut.
- managingEditor: Correu electrònic del responsable del contingut.
- webMaster: Correu electrònic del responsable tècnic.
- pubDate: Darrera data de publicació en el canal.
- category: Categoria del contingut.
- lastBuildDate: Darrera data de modificació del canal.
- generator: Programa fet servir per generar el contingut.
- docs: Descriu el format específic.
- ttl: Temps que els clients han d'esperar per tornar a demanar.
- image Icona que representa el canal.
- rating Fa servir una classificació americana sobre el contingut.
- cloud Grup de gent a la qual s'informa dels canvis.
- textInput És una etiqueta antiga que ja no es fa servir.
- skipHours En quines hores no es poden demanar actualitzacions.
- skipDays Quins dies no es poden demanar actualitzacions.

Contingut del canal

El contingut visible d'un canal RSS anirà en els elements <item>, dels quals pot haver la quantitat que es vulgui (fins i tot cap).

Normalment cada vegada que es produeix una novetat en el lloc associat al contingut d'un canal es crea un nou element item que s'afegeix al document.

A pesar que no sembla que tinga sentit crear ítems sense contingut, això és estrictament possible perquè ítem no té cap element obligatori però sempre hi ha d'haver un <description> o un <title>.

Per tant, el següent seria un document RSS correcte. Tenim dos ítems, un amb títol sense contingut i un amb contingut sense títol.

```
<rss version="2.0">
	<channel>
		<title>RSS</title>
			<link>http://ioc.xtec.cat/rss/RSS</link>
			<description>Provant</description>
		<item>
			<title>Títol sense contingut</title>
		</item>
		<item>
			<description>Contingut sense títol</description>
		</item>
	</channel>
</rss>
```

4. Atom

Atom va ser dissenyat pensant a superar els problemes d'interpretació que tenia RSS 2.0 i evitar la complexitat afegida d'RSS 1.0. La seva idea era aprofitar les millors coses dels RSS i arreglar les parts que en causaven confusió.

Un segon objectiu que el diferencia clarament d'RSS és que també es volia que no solament servira per recuperar els canvis en la informació del canal sinó que també es poguera fer servir de manera estandarditzada per afegir-hi informació. La sindicació ha estat molt lligada als blogs i fins al moment cada programa per fer blogs feia servir el seu protocol propi (Blogger API, MetaWebLog API, ...), que estaven pensats només per un blog en concret.

Per tant, podem dividir Atom en dues parts:

- Atom syndication format: un llenguatge XML per sindicar continguts.
- Atom publishing protocol: un protocol basat en HTTP pensat per actualitzar i crear recursos en el Web.

A part de les diferències en les etiquetes, les grans diferències amb RSS són que:

- Permet definir quin és el contingut de les etiquetes (text, HTML, etc.), però també permet referències a arxius externs. En RSS no es defineix quin contingut hi ha.
- Es pot fer servir dins d'altres documents XML, ja que té la seua definició i fa servir els espais de noms. No es pot posar RSS dins d'altres documents XML perquè no té en compte l'espai de noms.
- RSS no ofereix un protocol de publicació com Atom.

a. Llenguatge Atom

Atom té només una arrel, que és <feed>. Aquesta etiqueta la poden fer servir els programes per detectar que el document que estan llegint és de tipus Atom.

L'arrel <feed> sempre ha de tenir definit l'espai de noms dels documents Atom, que és www.w3.org/2005/Atom. Si no s'especifica l'espai de noms el document no validarà.

```
<feed xmlns="http://www.w3.org/2005/Atom">
      ...
</feed>
```

El fet de tenir un espai de noms i de fer-lo servir possibilita que els documents Atom es puguen mesclar amb documents XML d'altres vocabularis sense problemes.

Atom té disponibles els atributs d'XML xml:lang, que serveix per identificar l'idioma del document, i xml:base, que es fa servir per controlar com es resolen les adreces relatives.

De la mateixa manera que en altres llenguatges de canals, com RSS, es poden agrupar les etiquetes d'Atom en dos grups:

- Etiquetes que proporcionen dades sobre el canal.
- Etiquetes amb el contingut del canal.

Etiquetes amb dades del canal

Els elements obligatoris dins de l'etiqueta <feed> sempre han de tindre els elements fills <title>, <id> i <updated>

Per tant, aquest seria un document Atom vàlid, ja que aquestes són les úniques etiquetes obligatòries:

```
<?xml version="1.0" encoding="utf8"?>
<feed xmlns="http://www.w3.org/2005/Atom">
	<title>Atom IOC</title>
	<updated>2011-08-13T15:20:02Z</updated>
	<id>http://ioc.xtec.cat/</id>
</feed>
```

A part dels elements obligatoris, en l'especificació d'Atom també es fa referència a uns elements que anomena "molt recomanables" Els elements recomanables són <link> i <author>.

Etiquetes de contingut del canal

Per definir les diferents entrades dins d'un canal es fa servir com a base l'element <entry>. Cada nova aportació crearà un nou element <entry>, que com a mínim ha de tenir les etiquetes <title>, <id>, <updated>, i a més un element <content> o bé un element <link>

Com fa sovint l'especificació, Atom també defineix un segon nivell d'elements, considerats "molt recomanats", i que, per tant, també haurien de sortir. Entre els recomanats hi haurà <content> o <link> si no s'han especificat anteriorment.

L'element <author> es converteix en obligatori si no se n'ha especificat cap en les metadades del canal.

Format de les dates

Atom fa servir l'RFC 3339 (ISO 8601) per definir el format de les dates. Les dates en Atom han de tenir aquesta forma:

```
Any-Mes-DiaTHora:Minuts:Segons-zonahoraria
```

De manera que:

- Tots els valors són numèrics excepte la zona horària, que en alguns casos pot ser el caràcter "Z" per indicar l'hora universal.
- Davant de la zona horària s'especifiquen les hores de retard o d'avançament amb els símbols de suma o resta.
- Es fa servir la lletra "T" per separar els dies de les hores.

El contingut

Si no s'especifica cap atribut en l'element <content> o en <summary>, aquest serà tractat com si fos text pla. Si es vol deixar clar que el contingut és en algun altre format s'ha d'especificar amb l'atribut type, que normalment tindrà els valors "text", "html" o "xhtml".

5. Agregadors / lectors

Els agregadors i lectors de feeds són programes que permeten a l'usuari mantenir en un sol lloc tota la informació dels canals que li interessen.

Entre altres coses:

- S'encarreguen d'actualitzar els canvis que s'hi van produint sense que l'usuari hagi de visitar la pàgina.
- Porten el control dels continguts llegits i no llegits dels canals.
- Permeten veure un resum de les notícies d'un lloc web.
- S'hi poden organitzar les notícies en grups personalitzats.
- Permeten fer cerques d'informació entre la informació del canal.

Tant RSS com Atom són estàndards oberts, i això ha permès que s'hagin creat una gran quantitat de lectors que els suporten i que ofereixen funcions extra per als usuaris per intentar millorar-ne l'experiència.

A pesar que es poden llegir els canals des de diferents programes de correu o els navegadors, normalment s'aconsegueix treballar de manera més còmoda i personalitzable fent servir programes especialitzats.

En general podem dividir els programes lectors de canals en dos grans grups:

- Lectors web
- Lectors d'escriptori

A pesar de la divisió normalment l'aspecte visual d'aquests programes és relativament semblant. Tots solen tenir la pantalla dividida en blocs:

- En un dels blocs hi sol haver la llista de subscripcions en la qual es poden agrupar aquestes subscripcions per temes.
- Un bloc per mostrar el contingut de cada una de les subscripcions.

Exemple d'agregadors populars

Lectors via web	Programari d'escriptori
Feedly	Liferea
Netvibes	SharpReader
BlogLines	FeedDemon
FeedLooks	FeedReader

6. Bibliografia i referències

Sala, Xavier. (2023) Llenguatges de marques i sistemes de gestió d'informació.

Institut Obert de Catalunya