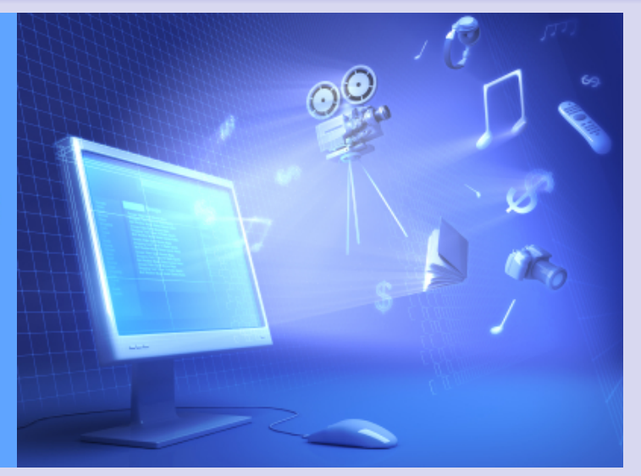
Archiving Ephemeral Data using Web Feeds

Marilena Oita and Pierre Senellart
InfRes department, DbWeb group, Telecom ParisTech
WebDam Project, Inria Saclay Ile-de-France

Archiving's aim: Preservation of Ephemeral Data

data going from factual to digital



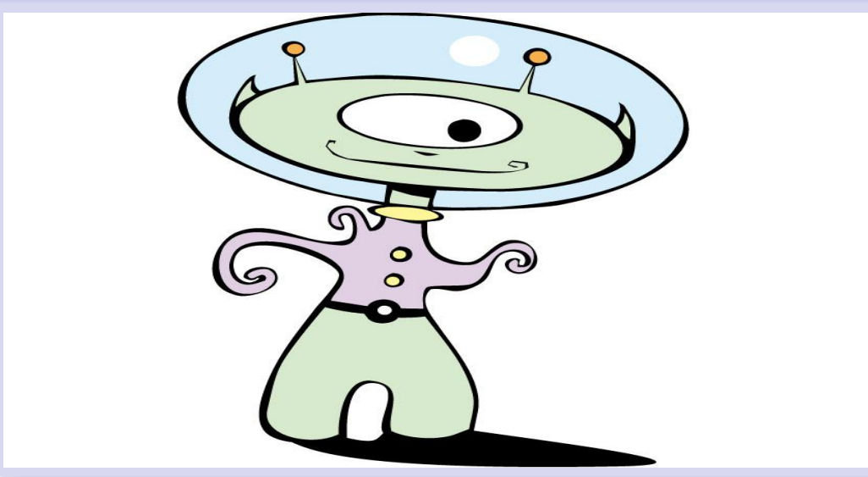
The Web's dynamics

a consequence of the Web 2.0 tools explosion

- frequently updated data
- new Web pages added each day

Information Retrieval from a (Personal) Web Archive

search Web data rooted in the past in a domain of interest



The Web Article's Extraction Technique

operating at DOM level: bottom-up strategy

1. use HtmlCleaner as a parser
2. filter the leaf nodes which contain at least one signifier as 'conceptual nodes'

Signifiers from the example on the right: *study*: concept, *being a scientist*: 3—gram.

Semantic Acquisition

extract signifiers from the feed item's title and description

- concepts: tokenize, stem and do a frequency analysis => a bag of relevant 'tags'
- n—grams: sequences of *n* words, taken as they appear in the title and description

Semantic Density Measure

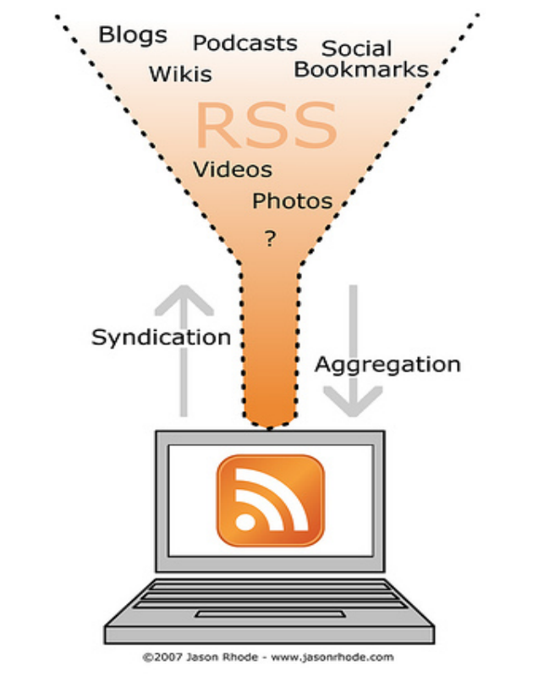
$$\textit{semanticDensity} = \sum_{n=1}^{nbConceptualNodes} \frac{cnode.nbOfSemanticMatches}{cnode.textualLength}$$

3. group the conceptual nodes in function of their lowest block-level common ancestor
4. take the block node which has the highest semantic density measure

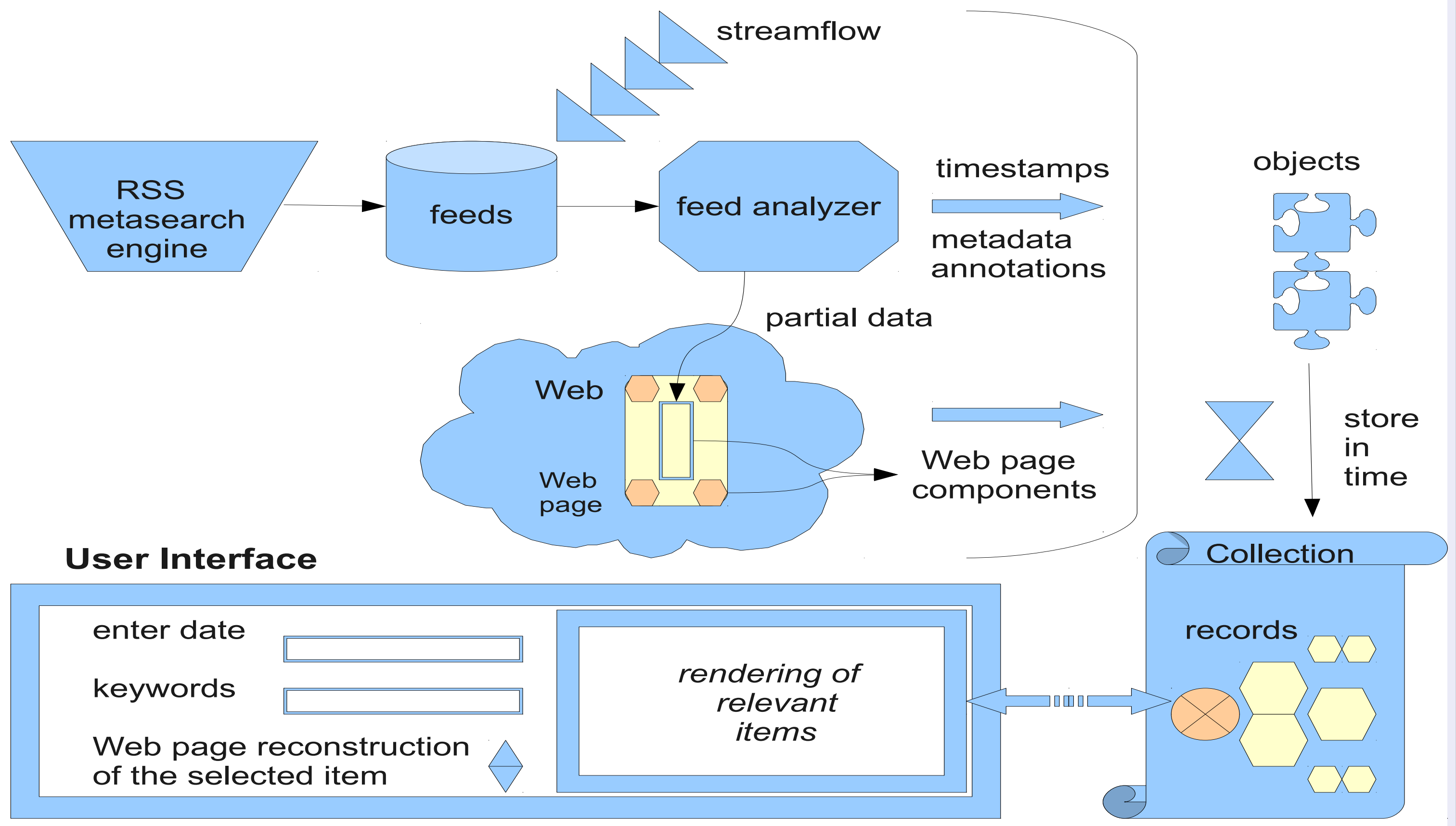
Web Feeds = XML-based files : RSS, Atom

crawl and analysis of domain-specific feeds

- pass through Search4RSS to acquire a list of feeds
- crawl the feeds rather than the Web pages
- do a semantic and temporal analysis using a feed parser



Uniformly Querying a Collection of Web Data Objects



Distinguishing between Semantic Zones

using concepts

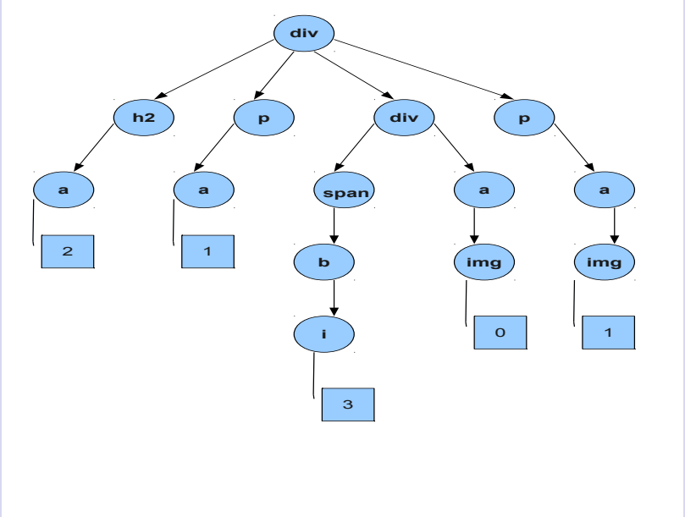
1. reconstruct the data object's context
2. identifies parts of the Web page that are semantically related to the item

using n-grams

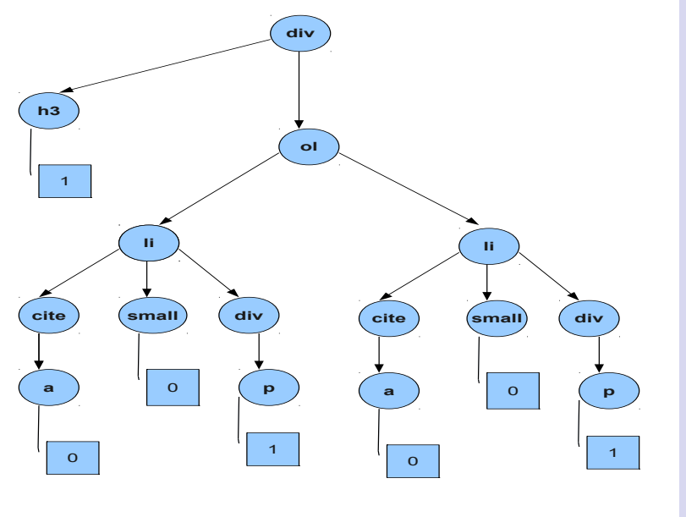
1. set the data object's content
2. increased precision to identify the zone that contains the article (if significant n—grams)

Example

article's node



comments'node



Heuristics:

1. the block-level node is a DIV
2. the comments' zone is encoded as a list

Web Feed leveraged Elements

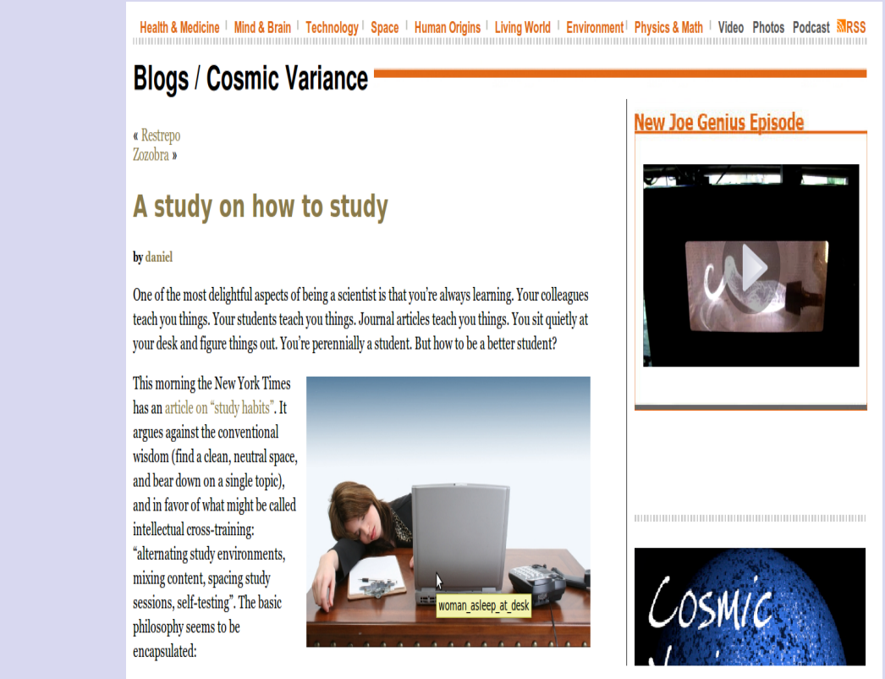
types of nodes

1. channel : the publication hub of a Web site
2. item: a resource uniquely identified by a URL and which has some semantics attached

important elements

- link
- title
- description
- pubDate: not compulsory, but still omnipresent

The Web page < ——



Blogs / Cosmic Variance

« Restrepo
Zozobra »

A study on how to study

by daniel

One of the most delightful aspects of being a scientist is that you're always learning. Your colleagues teach you things. Your students teach you things. Journal articles teach you things. You sit quietly at your desk and figure things out. You're perennially a student. But how to be a better student?

This morning the New York Times has an article on "study habits". It argues against the conventional wisdom (find a clean, neutral space, and bear down on a single topic), and in favor of what might be called intellectual cross-training: "alternating study environments, mixing content, spacing study sessions, self-testing". The basic philosophy seems to be encapsulated:

—— >The feed item

```
<item>
<title>A study on how to study </title>
<link>http://feedproxy.google.com/r̃/CosmicVarianceBlog/3̃/-
uatEVOIO0g/</link>
<comments>http://blogs.discovermagazine.com/
cosmicvariance/2010/09/07/a-study-on-how-to-study/#comments </comments>
<pubDate> Wed, 08 Sep 2010 03:16:54 +0000 </pubDate>
<guid isPermaLink="false"> http://blogs.discovermagazine.com/
cosmicvariance/?p=5353 </guid>
<description> <![CDATA[One of the most delightful aspects of being a scientist is
that you&#8217;re always learning.
Your colleagues teach you things. Your students teach you things. Journal articles
teach you things.
You sit quietly at your desk and figure things out. You&#8217;re perennially a
student.
But how to be a better student? This morning the New York [. . . ]]]>
</description>
<wfw:commentRss> http://blogs.discovermagazine.com/cosmicvariance/
2010/09/07/a-study-on-how-to-study/feed/ </wfw:commentRss>
<slash:comments>6 </slash:comments>
<feedburner:origLink> http://blogs.discovermagazine.com/cosmicvariance/
2010/09/07/a-study-on-ow-to-study/ </feedburner:origLink>
</item>
```

Web Object Signification and Components

- at feed level represents an item
- at Web page level represents a Web article

1. content: text and references
2. semantics: channel info (provenance), categories ('tags'), title
3. timestamps: the article's publication date and the date of crawl

Web Page Reconstruction

naturally excludes boilerplate

- extract and sort the semantic zones (in the analysis phase)
- keep them in a file
- download the .css files
- reconstruct the path to them (at run-time): domain-channelId-crawlTimestamp-itemId

Conclusions

contributions

1. Web feeds analysis
2. semantic data mining in the feed
3. a new way of extracting the relevant content of a Web page and the zones that are semantically related to it
4. storing information at data object level vs. at Web page level => smaller, cleaner versions