



What should be changed on the WWW to empower and lighten the operation of a search service ?

An opportunity for a new European project ?

Michel PLU - FRANCE TELECOM R&D Division

The present document contains information that remains the property of France Telecom. The recipient's acceptance of this document implies his or her acknowledgement of the confidential nature of its contents and his or her obligation not to reproduce, transmit to a third party, disclose or use for commercial purposes any of its contents whatsoever without France Telecom's prior written agreement.

Content search engine for FRANCE TELECOM



Electronic Contents: The next foreseen revenues for on line general public services

Q Music , video , Film , Images, books or any TV contents ...

Q Large audience perspective with large advertisement revenues

And personal contents sharing as a new form of communication services



alapage



Jukebox

France Telecom and Search Engine Research



- S A big R&D center with a goal of transferring new technologies into operational services and products
- S Unique telecom operator operating its own search engine technology
- S Clearly motivated to develop and integrate any technology to improve search quality efficiency and **COST**
- S Sensible that Research Projects takes into account operational and industrial issues
 - Q Not only rocket science ...

The current WWW



- S Facts:
 - Q 80% of WWW surfers navigations include an access to search service engine
 - Q Only 3 major players over the world : msn , yahoo , google
 - Even if locally some players can have a place (Wanadoo is 2nd in France)
 - Q Operating a search engine on the whole WWW requires huge investment and resources
 - Cf google 100000 computers

The current WWW



S Some remarks

- Q The WWW is no longer a distributed medium
 - A search engine is a centralized access point to WWW resources
 - For accessing a specific resource, typing URLs are being replaced by searching for keywords (cf google "I feel lucky")
- Q The current WWW is not efficient for being searched
 - Costs of search service are huge
- Q One suspects that the democratic, impartial, free qualities of the WWW are endangered if only few search engines operated by big commercial players can survive.
 - The WWW is viewed through search engine lens

And then , what more can we do ?



- S reduce cost of operating a search engine in order to see the development of multiple alternatives
 - Q if a search engine is a specific filtering process (an editorial choice) we can not accept having so few choices
- ⇒ Complement the Web in order to be more efficiently searchable

And then , what can we do ?



S Some examples detailed later :

Q New content management services embedded in the web

- Semantic DNS for content directories ?
- Certified logs management

Q New features in the protocols or new protocols dedicated to search engines

- Egs: think to: updated-since in http – URI managements for detecting copies or repurposed contents
- communication protocols dedicated to search engines

IMPLEMENTING FREE TOOLS SUPPORTING THOSE RECOMMENDATIONS RUNNING WITH EXISTING ONES

- Joins APACHE Foundation ?
- Operate a new group within W3C ?

Some examples of technical answers to be proposed



S Content discovery

Q Distributed semantic Content directory (a la DNS)

With several categories schemas but with mapping between them

Data validation filtering and ranking is an added value of services like search engine exploiting those data

Some examples of technical answers to be proposed



S Content crawling:

Q Reducing the huge overload and inefficiency of crawling contents

- Basic Simple solution for updates-awareness
 - *http server should support trustable management of: If modified since Content-length- unique digital signature of contents*

Q More efficient/sophisticated crawling protocols

- publishing can be more efficient than polling
 - *Active http servers sending notifications about content creations and important updates*
- Notifying only the interesting XML diff of updated content
 - *Cooperative http servers managing logs of new indexable contents and XML Diff of interesting updates*
- Transmitted in an efficiently encoded format (zip ?) containing only Information in a standardized format
- With a push model : Aggregating a set of new contents and updates (zipped XML diffs)

Some examples of technical answers to be proposed



S Distributed Content Indexation* with trusted partners :

Q Publishing efficient metadata

- Use local indexation provided by publishers
- Metadata can be automatically produced by local indexers or simply managed by Content Management Suite
- In order to be trustable local indexer and indexation process need to be certified **Would Develop the market of indexation tools**
- Metadata are provided to search engines services
 - *Eventually using published metadata Schema, and values and efficient URI management*

Q Metadata production for a content is only one step

- Final filtering and ranking of contents is another step managed by end users search engines

LARGELY RELEVANT FOR CONTENTS with ACCESS RIGHTS

*Metadata production

Some examples of technical answers to be proposed



S Relevance feedback for a better ranking

Q Why do we need to simulate a theoretical random walker model for PageRanking ?

Q Better exploit real user navigations **provided by each web sites**: Certified and **anonymous** – eventually with added information on user to build users or communities models

Q Using

- Certified logs manager
- Trusted / Certified audience measurement companies
- *Development of existing businesses*

Associated key issues to collaborate on



S Fighting against Spam indexing : a shared problem

Q Need shared resources

- If each search engines would share their blacklist then spamming becomes too risky compared to the profit
- *Being excluded from all the search engines*

Q Need regulation

- What if a spammer would loose its host name ?
- *Being excluded from the web*
- Need clear definition of spam behavior ..
- *A cyber-law ?*
- *A cyber court ?*

Associated key issues to collaborate on



S Content Access Rights

Q Are search engine robots allowed to access any content in order to index them ?

Q Are they allowed to only access metadata of contents (cf distributed content indexation) ?

Q How to filter contents accessible to users according to specific conditions ?
– Develop Metadata for access rights definition (MPEG 21 ?)

Associated key issues to collaborate on



S Plug and Play search engine architecture

Q Indexing technologies are evolving fast
– Many technology providers and a lot of research

Q New Contents to be indexed : broadcasted TV , podcast..

Q New queries from new devices: mobile phone , TV Set top box

=> For a search engine operator it is crucial to rapidly integrate
BEST available technologies with reduced COSTS

Q NEED of standardized APIs and DATA format

Q Component based architecture instead of a global monolithic search engine solution

Q Using efficient software technologies to integrate and activate adapted technology to requests and contents

And then , what can we do ?



S Is the answer only technical ?

S It is also about strategy and politics

QPartnerships

– Content producers, Content providers, Broadcasters , Search engines ...

– Finding win win strategies for all players in the content chain

QRegulation

QOrganization

Need to reach the critical mass to be credible and accepted

And then , what can we do ?



S An IP proposal With :

QContent producers, providers, broadcasters

QSearch engine and content management providers

QSearch engine and portals operators

QSEO (Search Engine Optimization) agencies

QAcademics specialized in Distributed information systems, Trust and reputation technologies, IP technologies, Multimedia Content indexation , Metadata management, Regulations, economics ...

And then , what can we do ?



S An IP proposal for validating ideas and exploring options to achieve a common and a shared goal:
Reducing the cost for supporting the operation of multiple alternatives of efficient world wide search engines

Qproducing leveraging tools :

*–mainly free software for supporting new protocols
recommendations or extensions between content providers and
operators of search engine services ...*