

RoSeS : Really Open, Simple and Efficient Syndication

ANR-07-MDCO-011-01

**Grand Colloque STIC
Lyon, 4-6 janvier 2011**

Projet RoSeS

ANR-07-MDCO-011-01



Début – fin :	1 jan 2008 – 31 déc 2011 (36 + 12 mois)
Site Web :	http://www-bd.lip6.fr/rosec
Intranet :	http://www-bd.lip6.fr/rosecprivate
Forge :	http://gforge.cnam.fr/gf/project/rosec/
Liste de Diffusion :	rosec@lip6.fr
Coût complet :	954 k€
<i>Aide allouée :</i>	294 k€
Partenaires :	4 partenaires académiques : • LIP6-UPMC (Wisdom) • CEDRIC-CNAM (Wisdom) • PRISM – UVSQ • LSIS – Univ. de Toulon 1 partenaire industriel : • 2or3things

Contexte : Web 2.0 et RSS



- Ressources Web 2.0 :
 - évolutives : actualités, blogs, ...
 - accessibilité et utilité limitées dans le temps
 - partagées avec des communautés
- Syndication web :
 - faciliter la **diffusion ciblée et l'agrégation** d'informations sur le web
- Trois principes :
 - flux d'items XML : titre, date, auteur, description, liens
 - principe de « publication/souscription »
 - formats standards : RSS et ATOM



Missions et Organisation

Programmes de recherche

Partenariats Public Privé

Magazine

Recherche

Ok

Programmes de recherche > Appels à projets

Programmes de recherche

Appels à projets

Calendrier 2012

Actions de diffusion de la culture scientifique

Projets sélectionnés

Projets financés

Documents

Charte de déontologie

Règlement financier

Questions fréquentes

Départements

Comités

Investissements d'avenir

Ateliers de réflexion prospective

Guide de l'appel à projets

Appels à projets

Par cette rubrique vous accédez à l'ensemble de nos appels à projets. Vous pouvez effectuer votre propre sélection en affichant les appels à projets selon leur état, ou uniquement ceux d'un programme en particulier, ou encore faire une recherche libre par mots-clés.

Calendrier prévisionnel de lancement des appels à projets 2012

Présentation des appels à projets 2012

En parallèle à ses appels à projets, l'ANR publie des appels à propositions dans le cadre d'Ateliers de réflexion prospective.

Thèmes

Mots-clé

Année de programmation

Etat

☐ Clos
☐ Financement
☒ Ouvert
☐ Sélectionné

Actualiser

Résultats de votre recherche

22 résultats correspondent à votre recherche

Thème	Nom	Etat	Date de clôture
Biologie - santé	Appel à projets franco-californien de recherche fondamentale sur les cellules souches		10/01/2012
Biologie - santé	Appel à Projets Transnational pour la recherche sur les maladies rares, conduite par des jeunes chercheurs (dans le cadre de l'ERA-NET E-Rare-2)		31/01/2012
Biologie - santé	Technologies pour la santé et l'autonomie (TecSan)		01/03/2012
Biologie - santé	Maladie d'Alzheimer et Maladies Apparentées (MALZ)		14/02/2012
Energie Durable	Transports Durables et Mobilité		07/03/2012

Colloques

Du 04/01/2012 au 06/01/2012

Grand colloque bilan STIC, Lyon

Du 10/01/2012 au 10/01/2012

Colloque "SERVICES CLIMATIQUES : vers la fourniture d'information climatique pour le secteur industriel" - Paris

Du 12/01/2012 au 13/01/2012

Quelle recherche pour les Energies du futur ? 2ème Colloque Bilan Énergie

Toute l'Actu

2012

"Le goût des sciences 2011" pour le projet ANR JCJC 2009 : DEFORMATION (étude de la dynamique de bulles d'air dans un fluide)

Inscrivez-vous à notre prochain colloque "Sciences et Technologies de l'information et de la Communication"

WISG12 : Sixième édition de la conférence sur la recherche en Sécurité Globale

Colloque bilan et perspectives "Santé-environnement et santé-travail", 1er et 2 février à Paris

ANR

212, rue de Bercy

75012 Paris

01 78 09 80 00

Hotline Appels à projets : 0811 42 8000

WISDOM

Grand C

4

[Accueil](#)
[Profil](#)
[Messages](#)
[Suggestions](#)

Cap Digital

@Cap_Digital Paris
The French Business Cluster for Digital Content
<http://www.capdigital.com>

[+ Suivre](#)

Tweetez à @Cap_Digital

Tweets Favoris Abonnements Abonnés Listes

Cap_Digital Cap Digital
Investissement d'avenir-FSN-AAP R&D e-Education n°2 est ouvert !: [19 mars 2012;] Lancement de l'appel e-Edu... bit.ly/uMopcf
Il y a 4 heures

Cap_Digital Cap Digital
Congrès National de la valorisation des contenus numériques: [3 janvier 2012;] Les 25 et 26 janvier 2012, à l... bit.ly/tLzjzC
Il y a 6 heures

Cap_Digital Cap Digital
IDIFF 2012, 10 ans déjà !: [2 janvier 2012;] 17 & 18 janvier 2012 au Palais des ... bit.ly/tu5oxy
2 Janv

Cap_Digital Cap Digital
Invitation – Lancement du groupement d'employeurs: [25 janvier 2012;] Inscription bit.ly/s2OWgg
2 Janv

Cap_Digital Cap Digital
Appel à projets EXAPAD: [16 janvier 2012;] Le Département de Paris et ses partenaires ont lancé un appel à p... bit.ly/uFQqB8
2 Janv

A propos de @Cap_Digital

1 038 Tweets 340 Abonnements 2 905 Abonnés 22 Lis

Images récentes · voir tout

Vous et @Cap_Digital
Vous suivez un compte abonné à @Cap_Di

Semblable à @Cap_Digital · voir t

Capucine_Cousin Capucine Cousi
Journaliste à Stratégies, tech, nux m

rafigaro rafi Haladjian · Suivre
I wear glasses, I hate cucumbers, I

fano Stéphane Distinguin · Suivre
Je cherche encore ! En attendant : fo

Abonnements · voir tout

À propos Aide Blog Mobile Statut Offres d'e
Confidentialité Raccourcis Annonceurs Profes
Développeurs Ressources © 2011 Twitter



flickr de YAHOO!

Accueil La visite Inscription Explorer Importer

Vous n'êtes pas connecté Connexion Aide

Rechercher

Galerie de photos de CapDigital

Classes Albums Expos Tags Personnes Archives Favoris Profil



Visiteurs

Les rencontres ont été consacrées aux étapes clés de l'année 2010, à l'Institut de la Vie Numérique,...

Tous droits réservés
Ajoutée le 18 avr 2011
0 commentaire



Cocktail d'échanges entre adhérents

Les rencontres ont été consacrées aux étapes clés de l'année 2010, à l'Institut de la Vie Numérique,...

Tous droits réservés
Ajoutée le 18 avr 2011
0 commentaire



Rencontres à la Bourse de Commerce

Les rencontres ont été consacrées aux étapes clés de l'année 2010, à l'Institut de la Vie Numérique,...

Tous droits réservés



Présentation des Projets R&D de Cap Digital

Les rencontres ont été consacrées aux étapes clés de l'année 2010, à l'Institut de la Vie Numérique,...

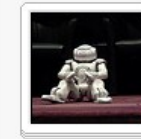
Tous droits réservés

Diaporama Partager



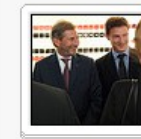
Rencontres Cap Digital...

7 photos
21 vues



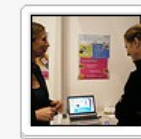
Future Game On 2010

15 photos
13 vues



Cap Digital - Visite du...

15 photos
17 vues



Cap Digital - Nathalie...

14 photos
27 vues



Cap Digital - Les vœux de...

4 photos
11 vues

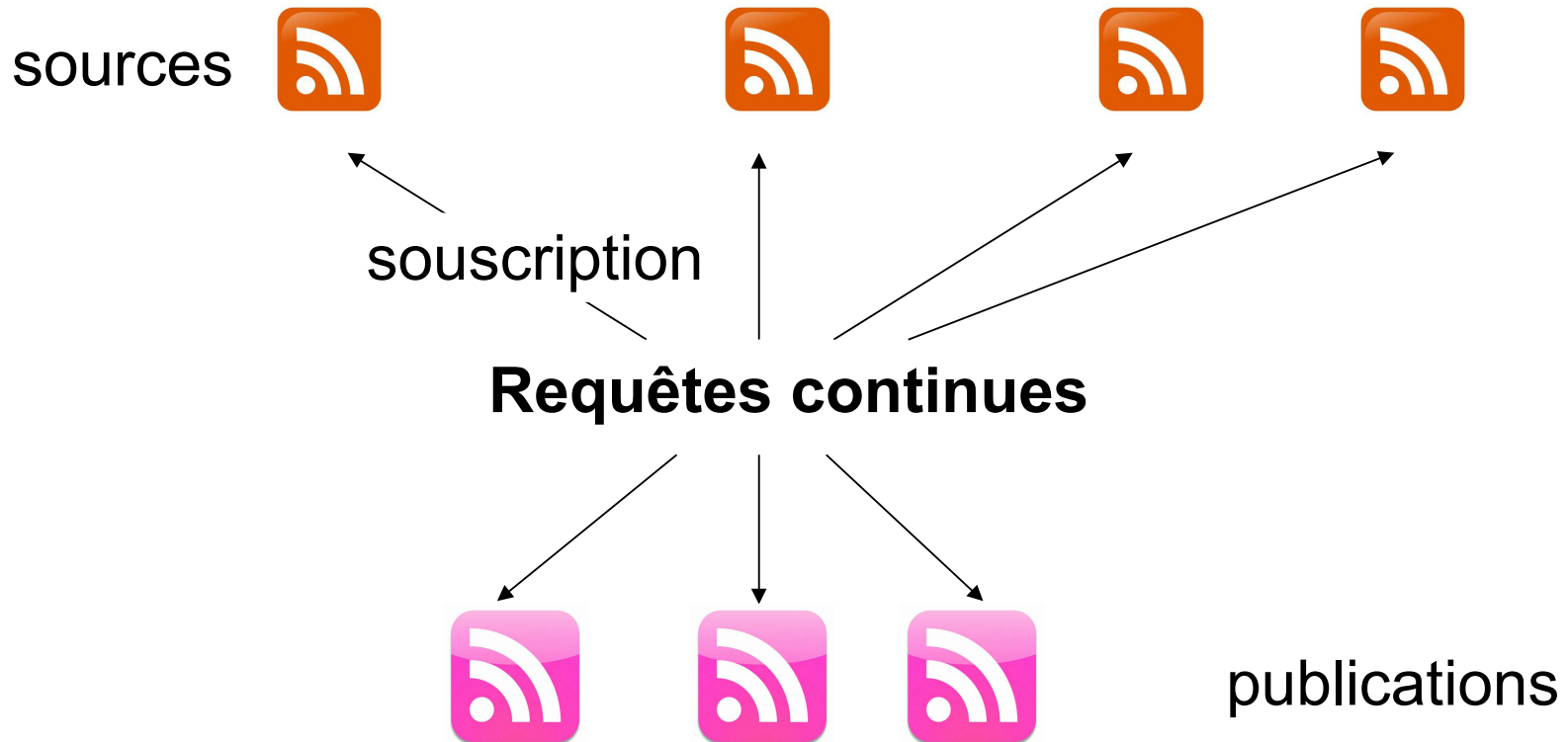
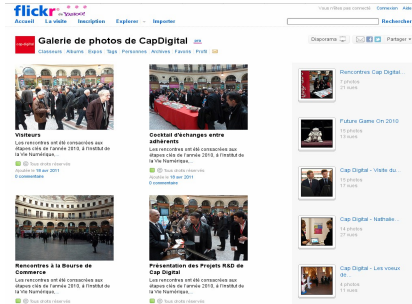
Agrégateurs RSS



Agrégateurs de flux RSS				
	Filtrage	Classification	Annotation	Recommandation / communauté
Google Actualités	mots-clés	x	-	-
Yahoo! Actualités	mots-clés	x	-	-
Google Reader	mots-clés	x	x	item / google friends
Technorati	mots-clés	x	x	source
Reddit	mots-clés	x	x	vote / commentaire
Digg	mots-clés	x	x	click / facebook
Newsvine	mots-clés	x	x	commentaire

- Opérations / langages limitées (mots, clés) et / ou spécialisés
- Traitements implicites / personnalisation limitée
- Pas de réelle composition de flux (par exemple jointure)

Agrégateur RoSeS



Approche et problèmes



Requêtes continues
Pub-Sub

Flux de données



XML

Passage à l'échelle

Modélisation
Langages de requêtes

Optimisation
Distribution

Workpackages



Modélisation et Langages (WP1 et WP5)

Architectures et Traitements (WP1 et WP3)

Distribution et optimisation (WP4)

Validation (WP6)



Modèle RoSeS



- Flux de fragments XML simples
- Langage de syndication
- Algèbre logique : filtrage, fenêtrage, jointure
- Modèle et langage logique « pivot » décliné dans différents contextes applicatifs et approches de traitement
 - Agrégation RSS
 - Analyse de séries temporelles (données boursières)
 - Surveillance de réseaux de capteurs

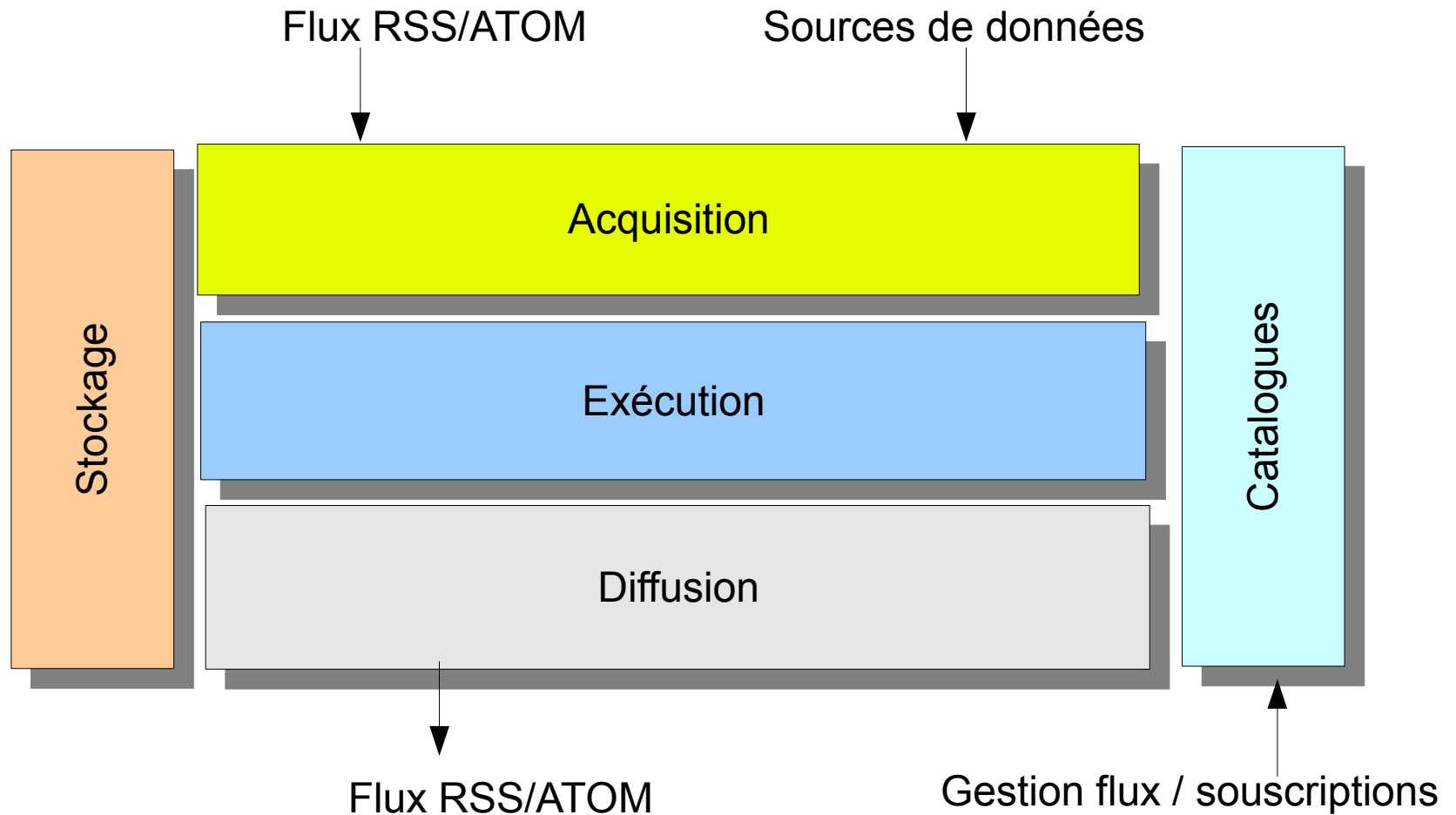


Langage de syndication

- Enregistrement : source → RoSeS
 - register source **google** as GoogleAlert('football') until '01/09/2010'
 - register source **myplayers** as document('MyPlayers.xml')/player'
- Publication : RoSeS → RoSeS
 - publish channel **football** as for \$i in **google**, \$p in **myplayers**
where \$i contains \$p/name return \$i
 - publish channel **euro2008** as for \$i in **football** where \$i contains « Euro 2008 » return \$i
- Souscription : RoSeS → destination
 - subscribe to channel **football** where \$i contains « Ronaldo » format RSS at most 10 items
 - subscribe to channel **Ronaldo** format RoSeS



Architecture fonctionnelle



Deux Approches de Traitement



- Approche « requête continue » (LIP6, CNAM):
 - Évaluation en continue (sans stockage de données)
 - Algèbre physique et plan d'exécution
 - Applications : filtrage d'actualités, médias sociaux
- Approche « XQuery + trigger » (LSIS, Prism)
 - Données stockées dans un entrepôts XML
 - Requêtes XML déclenchées par des événements (triggers)
 - Applications : analyse de séries temporelles (flux boursiers), agrégation avec données statiques, archivage

Problèmes Fondamentaux



- Acquisition de flux RSS (crawling)
- Stockage et interrogation P2P de séries temporelles
- Indexation de requêtes XML continues
- Indexation de souscriptions textuels
- Optimisation multi-requêtes
- Génération de bancs d'essais

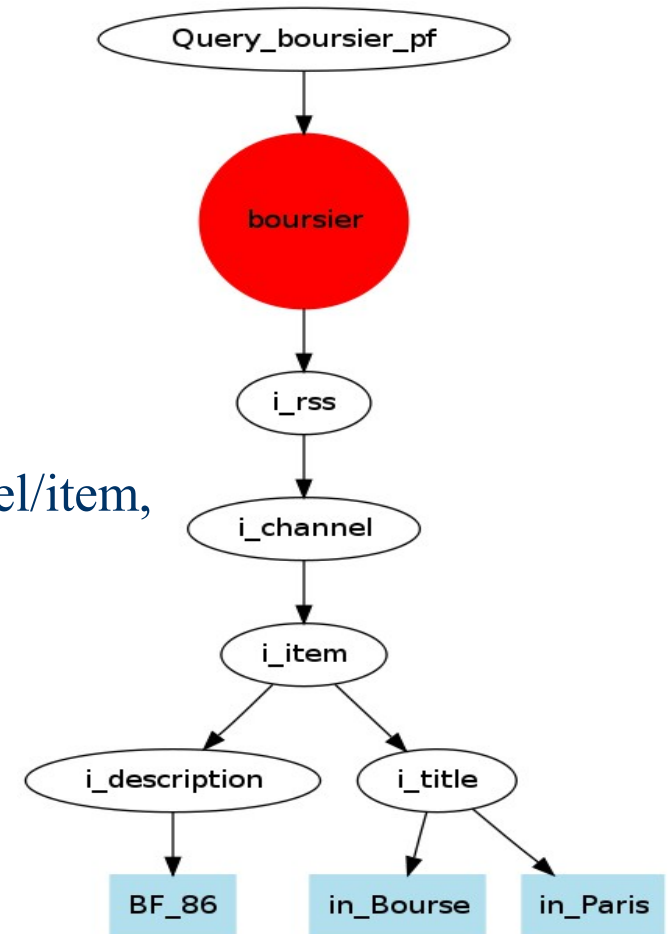
Indexation de requêtes XML (PRISM)



- Requêtes XQuery continues
- Indexation de Motifs Arbres à Prédicats (MAP)

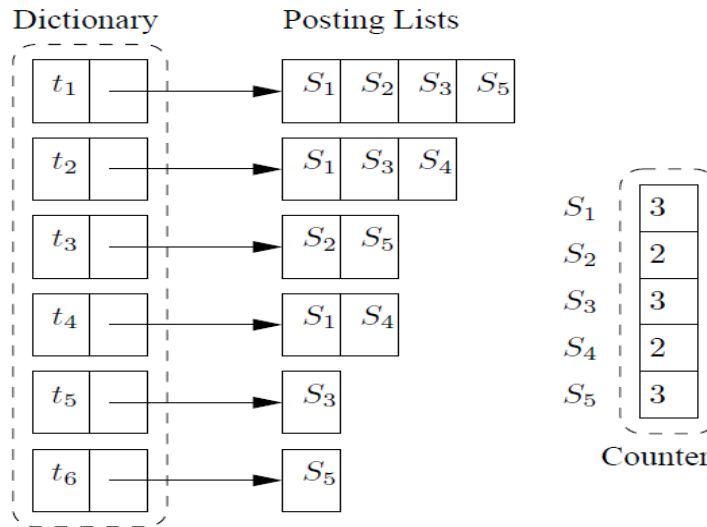
```

for $i in feed(tumbling, 1, new,'boursier')/rss/channel/item,
    $pf in /data/portefeuille
where contains($i/title,'Bourse')
    and contains($i/title,'Paris')
    and contains($i/description, $pf/action/nom)
return <a_lire>{$i}</a_lire>
  
```

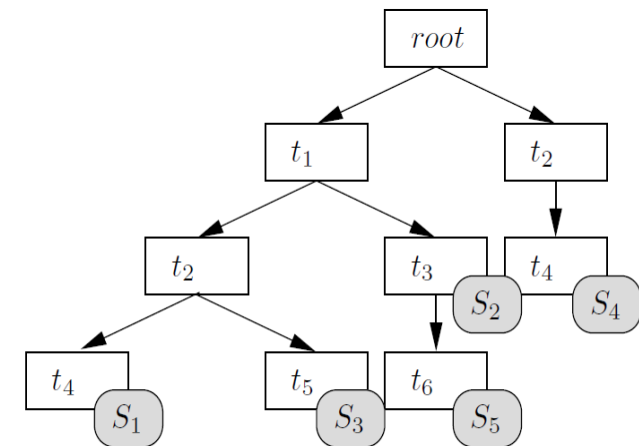


Indexation de souscriptions (CNAM)

- Souscription : ensemble de mots
- Listes inversées :
 - Count-based, Ranked-key
- Arborescences :
 - Regular Ordered Trie, Patricia Ordered Trie

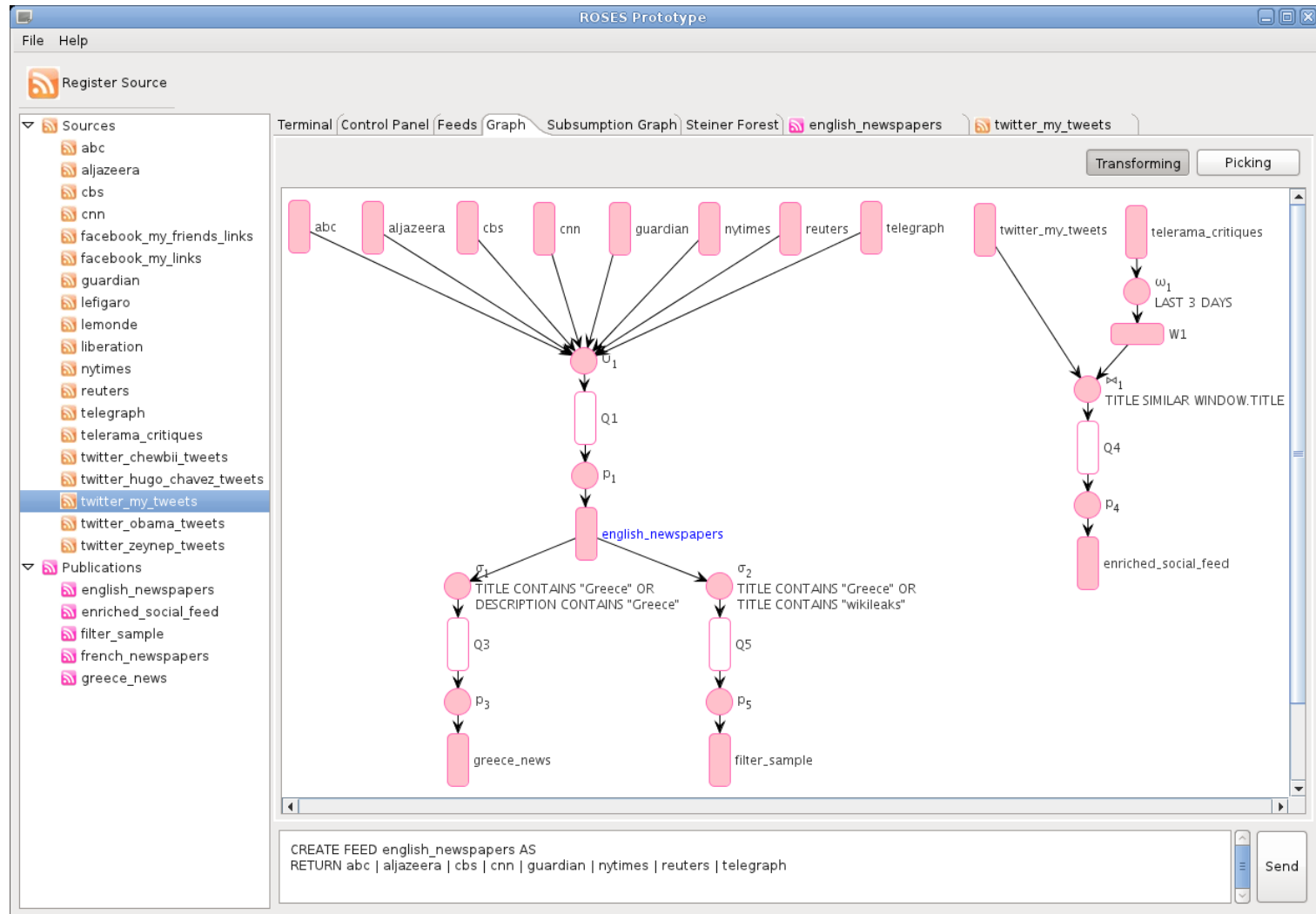


(a) Counted-based inverted list



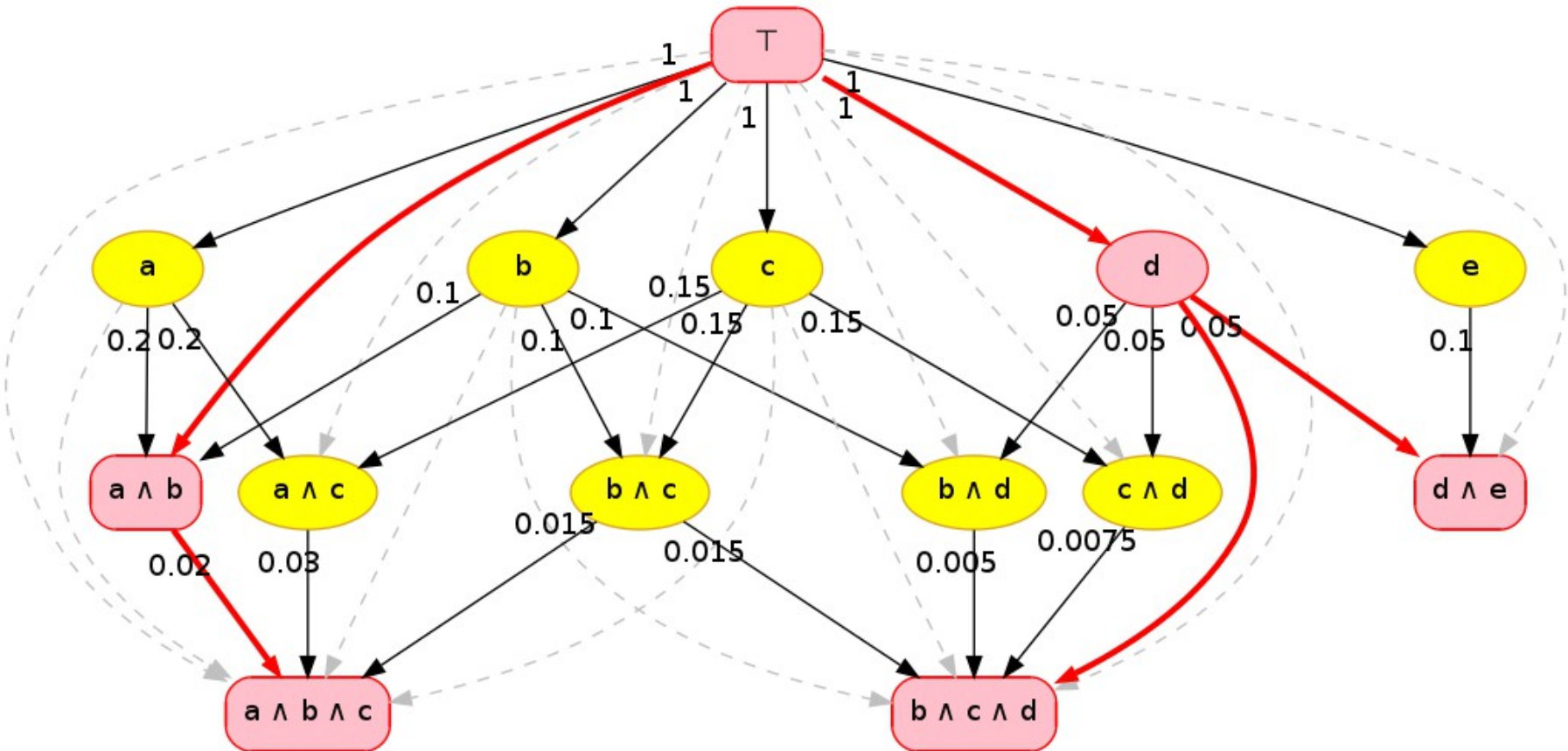
(c) Order-trie index

Optimisation multi-requêtes (LIP6/CNAM)



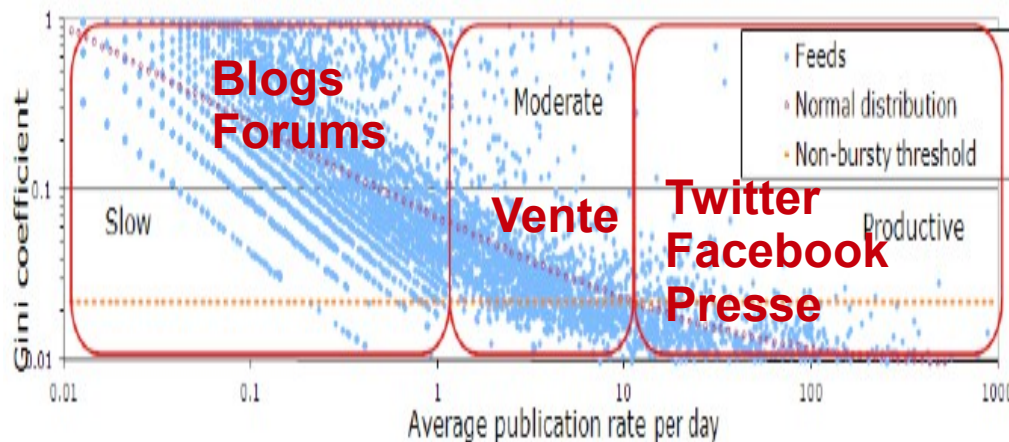
Plans de filtrage complexes

Arbre de Steiner minimal



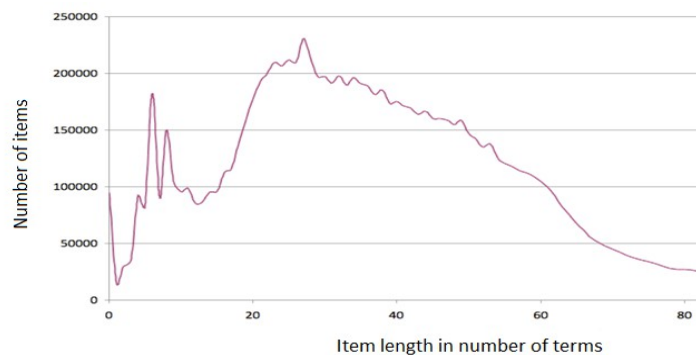
Analyse du web RSS (CNAM)

Fréquence de publication



Type	% feeds	% items	ratio
Social Media	1.77%	9.45%	7085
Press	9.99%	38.82%	5141
Forum	1.51%	3.62%	3178
Sales	11.32%	15.49%	1811
Misc.	41.47%	25.47%	812
Blog	33.93%	7.14%	278

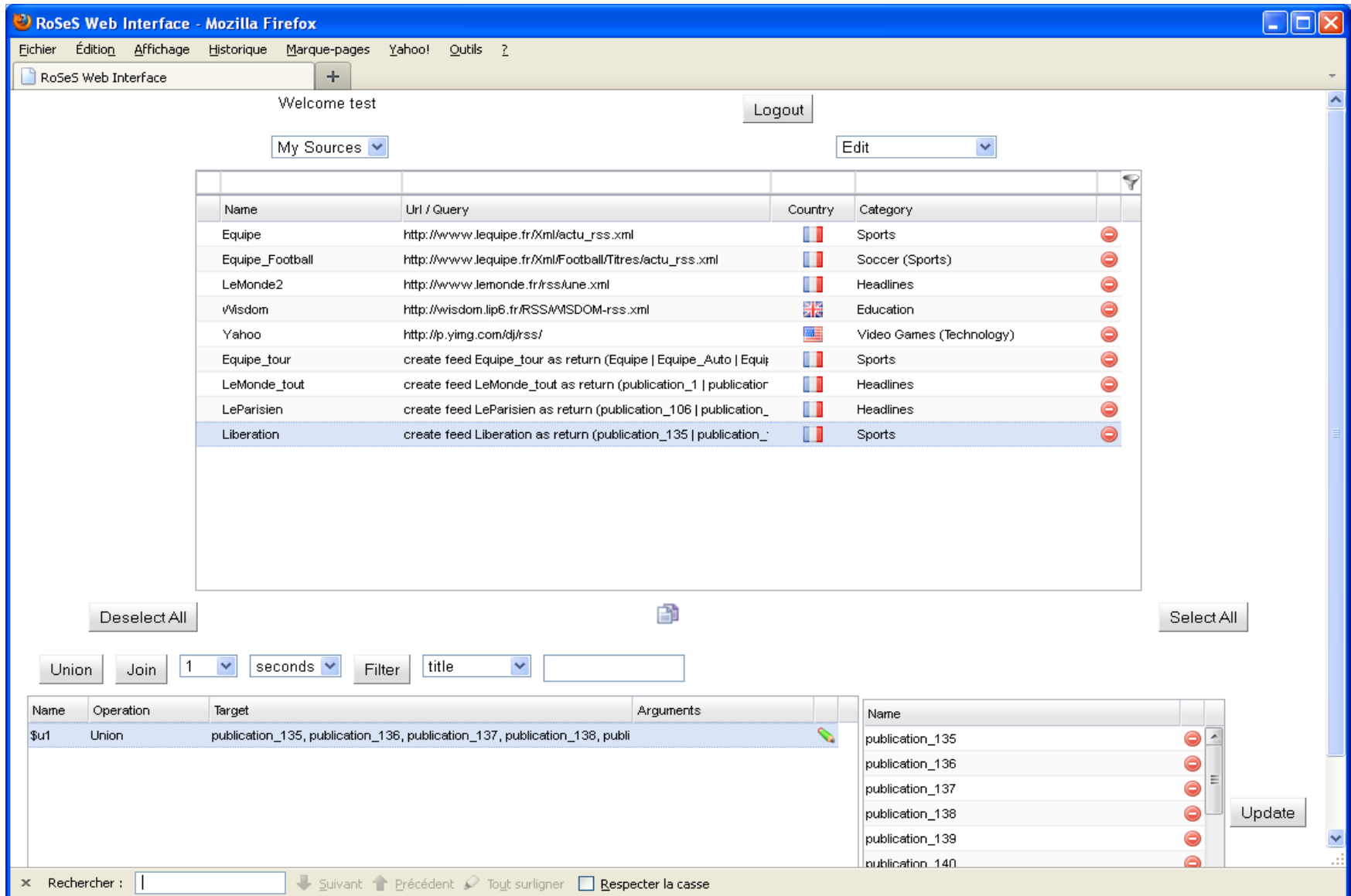
Longueur items



Contenu

- Termes : occurrences, co-occurrence
- Distribution exponentielle étirée
- Vocabulaire / taille
- Loi de Heaps (standard)

Prototypes



RoSeS Web Interface - Mozilla Firefox

Echier Édition Affichage Historique Marque-pages Yahoo! Outils ?

RoSeS Web Interface +

Welcome test Logout

My Sources Edit

Name	Url / Query	Country	Category
Equipe	http://www.lequipe.fr/Xml/actu_rss.xml	FR	Sports
Equipe_Football	http://www.lequipe.fr/Xml/Football/Titres/actu_rss.xml	FR	Soccer (Sports)
LeMonde2	http://www.lemonde.fr/rss/une.xml	FR	Headlines
vWisdom	http://wisdom.lip6.fr/RSS/WISDOM-rss.xml	GB	Education
Yahoo	http://p.yimg.com/dj/rss/	US	Video Games (Technology)
Equipe_tour	create feed Equipe_tour as return (Equipe Equipe_Auto Equi	FR	Sports
LeMonde_tout	create feed LeMonde_tout as return (publication_1 publication	FR	Headlines
LeParisien	create feed LeParisien as return (publication_106 publication_	FR	Headlines
Liberation	create feed Liberation as return (publication_135 publication_	FR	Sports

Deselect All Select All

Union Join 1 seconds Filter title

Name	Operation	Target	Arguments
\$u1	Union	publication_135, publication_136, publication_137, publication_138, publi	

Name
publication_135
publication_136
publication_137
publication_138
publication_139
publication_140

Update

Rechercher : Suivant Précédent Tout surligner Respecter la casse



Conclusion : Résultats et Bilan



- Contributions scientifiques :
 - Modélisation, stockage et interrogation de flux RSS
 - Nouvelles techniques d'optimisation de requêtes continues
 - Analyse statistique du web RSS
- Logiciels développés :
 - *3 prototypes complets fondés sur différentes architectures et contextes (voir démonstrations)*
 - *Générateur de banc d'essais avec des simulateurs*
- Encadrement et publications :
 - 1 thèse (financée par le projet) et 6 autres thèses qui ont partiellement contribué au projet
 - 7 stages (essentiellement M2)
 - 11 publications directement liées au projet
- Collaborations avec FORTH/Univ. de Crète et Univ. d'Arizona

Conclusion : Perspectives

- Requêtes continues pour la personnalisation de flux de données
 - Journalisme de données, open-data
 - Twitter (thèse en cours au CNAM)
 - Requêtes top-k sur les actualités (thèse en cours au LIP6)
 - Mobilité et données ambiantes (PRISM)
 - Personnalisation de la surveillance de capteurs (LSIS)

WISDOM

LIP6

vertigo

6

LABORATOIRE
D'INFORMATIQUE
PRISM

2or3things

ANR
AGENCE
NATIONALE
DE LA
RECHERCHE

cap·digital



Merci pour votre attention

- Site web : <http://www-bd.lip6.fr/rozes>
- Forge : <http://gforge.cnam.fr/gf/project/rozes/>



Publications



1. Characterizing Web Syndication Behavior and Content. Zeinab Hmedeh, Nicolas Travers, Nelly Vouzoukidou, Vassilis Christophides, Cedric du Mouza, Michel Scholl, WISE'11, The 11th International Conference on Web Information System Engineering, October 2011, pp.29–42, Series LNCS, Sydney, Australia
2. Everything you would like to know about RSS feeds and you are afraid to ask. Zeinab Hmedeh, Nicolas Travers, Nelly Vouzoukidou, Vassilis Christophides, Cedric du Mouza, Michel Scholl, BDA'11, Base de Données Avancées, October 2011, Rabat, Maroc
3. RoSeS: A Continuous Query Processor for Large-scale RSS Filtering and Aggregation. Jordi Creus Tomàs, Bernd Amann, Nicolas Travers, Cristian Dan Vodislav, démonstration, CIKM 2011
4. RoSeS: A continuous content-based query engine for RSS feeds, Jordi Creus Tomàs, Bernd Amann, Nicolas Travers, Dan Vodislav, 22nd International Conference on Database and Expert Systems Applications (DEXA 2011), Toulouse, France, August 2011.
5. Optimizing large collections of continuous content-based RSS aggregation queries, Jordi Creus Tomàs, Bernd Amann, Vassilis Christophides, Dan Vodislav, Nicolas Travers, BDA 2011
6. Online Refresh Strategies for RSS Feed Crawlers, Roxana Horincar, Bernd Amann, Thierry Artières, BDA 2011
7. Best-effort refresh strategies for content-based RSS feed aggregation, R. Horincar, B. Amann and T. Artières, The 11th International Conference on Web Information System Engineering (WISE 2010), Hong Kong, China, December 12-14, 2010
8. RoSeS : Un agrégateur de flux RSS avancé, J. Creus Tomas, N. Travers, B. Amann, D. Vodislav, Bases de Données Avancées (BDA 2010), Démonstration, Toulouse, France, October 2010.
9. A Semantic Map of RSS Feeds to support Discovery, Gaiane Hochard, Zoé Lacroix, Jordi Creus et Bernd Amann, 3rd International Workshop on REsource Discovery, November 5, 2010 Paris, France, Europe (workshop joint à 12th International Conference on Information Integration and Web-based Applications & Services - iiWAS2100)
10. Efficient P2P Processing of Times Series: Application to Stock Investment and Mobile Objects Analysis. G. Gardarin, B. Nguyen, L. Yeh., et.al., BDA'09
11. XQ2P: Efficient XQuery P2P Time Series Processing, B. Butnaru, B. Nguyen, G. Gardarin, L. Yeh. , (démonstration)

Logiciels développés



- Crawler de flux RSS/ATOM
- Moteur de requêtes continues (*démonstration*)
- Moteur XQuery avec cache P2P pour séries temporelles
- Moteur XQuery + trigger avec indexation de motifs d'arbres (*démonstration*)
- Moteur XQuery + trigger embarqué (*démonstration*)
- Simulateur de flux RSS
- Générateur de souscriptions RSS



Stages



1. S. Attrash. Moteur de requêtes continues RSS, stage M2, CEDRIC/LIP6, 2008
2. F. Feugeas, Q. Mansuy. Mise en œuvre des couches basses d'un mini serveur pour les pairs du réseau et adaptation d'un réseau P2P pour l'indexation de signatures de requêtes XQuery sur des flux RSS, PRISM, 2009.
3. I. Mami. Optimisation de requêtes continues, stage M2, CEDRIC/LIP6, 2009
4. M. Diouri, Indexation pour la recherche par le contenu textuel de flux, stage M1, CEDRIC, 2009
5. T. Dailly, Réalisation d'un simulateur de flux Atom en J2EE à partir des données de Wikipedia, LSIS, 2009
6. F. Troïlo, Réalisation d'un monteur d'acquisition et de persistance de flux RSS en J2EE, LSIS, 2009
7. N. Vouzoukidou, Générateur de souscriptions RSS, CEDRIC/FORTH, 2011

Thèses



1. **Jordi Creus (LIP6 / CEDRIC) : Modélisation et traitement continue de flux RoSeS (thèse financée par le projet)**
2. Bogdan Butnaru (PRiSM) : P2P et séries temporelles
3. Iulian Sandu-Popan (PRiSM) : Interrogation de séries temporelles
4. Roxana Horincar (LIP6) : Synchronisation de flux RoSeS
5. Zeinab Hmedeh (CEDRIC) : Indexation de souscriptions RSS
6. Ryadh Dahimen (CEDRIC) : Filtrage avancé de flux Twitter
7. Nelly Vouzoukidou (LIP6) : Requêtes top-k sur des flux de news