

RoSeS : Really Open, Simple and Efficient Syndication

Revue mi-parcours
2 septembre 2009
ANR-07-MDCO-011-01

Plan

- Contexte et objectifs du projet
- Activités scientifiques et techniques
- Fonctionnement du consortium
- Dissémination et valorisation



RoSes



Contexte et objectifs

Contexte : Syndication web

- Objectif : diffusion *efficace* et *ciblée* de flux d'informations sur le web
- Principes de syndication :
 - *Publication* : Les fournisseurs / sites web publient des canaux qui diffusent des *flux* d'informations
 - *Souscription* : Les consommateurs / clients intéressés se souscrivent à ces canaux pour avoir les dernières informations *en continu*.
 - *Agrégation et syndication* : Les flux publiés peuvent être agrégés et republiés par des tiers.
- En pratique :
 - Deux formats XML : RSS et ATOM 
 - Un grand nombre de sources et d'applications

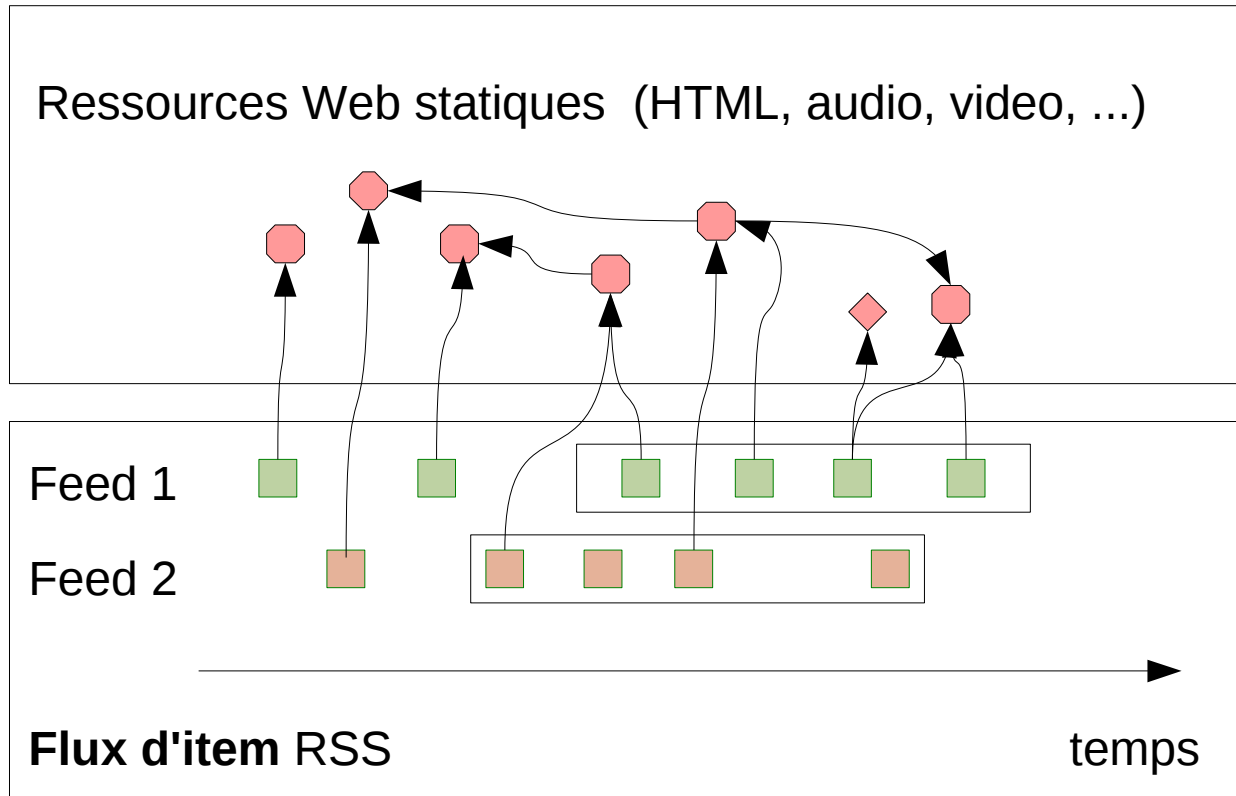


RSS Exemple

<http://www.agence-nationale-recherche.fr/anrinfos.php> :

```
<?xml version="1.0" encoding="UTF-8"?>
<rss version="2.0">
  <channel>
    <title>ANR</title>
    <link>http://www.agence-nationale-recherche.fr/</link>
    <description>Actualités de l'Agence Nationale de la Recherche</description>
    <item>
      <title>Projets sélectionnés HPAC, RiskNat et Blanc international</title>
      <link>http://www.agence-nationale-recherche.fr/Select2009</link>
      <description>Les listes de projets sélectionnés "Hydrogène et Piles à Combustible" et
        "Risques Naturels : Compréhension et Maîtrise", ainsi que les listes de projets
        franco-japonais et franco-taïwanais sont en ligne
      </description>
      <pubDate>Mon, 20 Jul 2009 16:06:20 +0200</pubDate>
      <guid>http://www.agence-nationale-recherche.fr/Select2009</guid>
    </item>
    <item>...</item>
  </channel>
</rss>
```

Syndication RSS



- Formats hétérogènes et complexes
- Mode « pull »
- Liens de navigation

- Formats homogène et simple
- Mode «push »
- Liens d'annotation

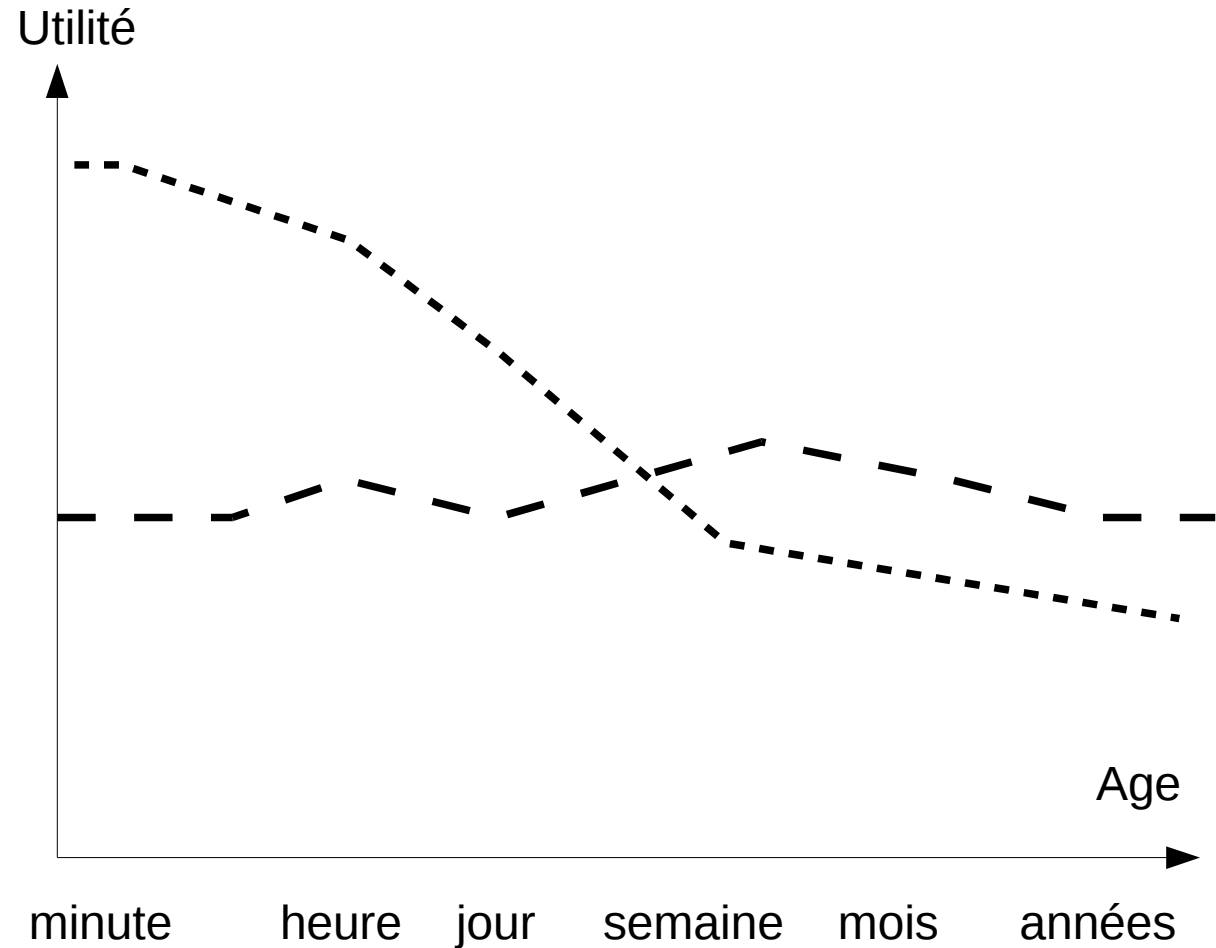
Contexte : Le Web « RSS »

Item RSS :

- changements
- résumé simple
- utilité décroissante

Page web :

- information(s)
- contenu riche
- utilité constante



Le Web « RSS »

Générateurs de flux RSS	
Journaux et agences de presse	Le Monde, Reuters, Libération, AFP, ...
Podcast	Audio : Radios, iTunes Store
	Vidéo : Youtube, Dailymotion, Google Video
	Images : Flickr, Google Images
Réseaux sociaux	Twitter, Facebook
Blog (Blogosphères)	Google Blogger, Windows Live Space
Autres sites web	...

- | | | |
|---|---|--|
| <ul style="list-style-type: none"> Grand nombre de canaux avec 10 à 100 items / par jour Items / messages complémentaires et redondants | → | <ul style="list-style-type: none"> Filtrage Classification Annotation Recommandation |
|---|---|--|

Agrégateurs RSS

Agrégateurs de flux RSS				
Outils	Recherche	Classification	Annotation	Recommandation / communauté
Google	mots-clés	x	-	-
Yahoo!	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

- Observations :
 - Annotation / recommandation *manuel*
 - Pas de *composition de flux* (jointure)

Yahoo Pipes

- Portail avec un éditeur graphique pour la création de requêtes d'agrégation
- Agrégation = graphe de composition d'opérations
 - Sources diverses / adaptateurs spécialisés
 - Opérations : filtrage, projection, union, jointure, tri, ...
 - Yahoo Query Language (début 2009)

Yahoo Pipes

le monde*

Layout Expand All Collapse All

Sources

User inputs

Operators

- Count +
- Create RSS +
- Filter +
- Location Extractor +
- Loop +
- Regex +
- Rename +
- Reverse +
- Sort +
- Split +
- Sub-element +
- Tail +
- Truncate +
- Union +
- Unique +
- Web Service +

Url

String

Date

Location

Number

Favorites

My pipes

Fetch Feed

- + URL
- <http://www.lemonde.fr/rss/une.xml>

Filter

Permit items that match any of the following

- + Rules
- item.title Contains text [wired]

Put search string (text)

Name: filter

Prompt: Put search string

Position: number

Default: paris

Debug: paris

Pipe Output

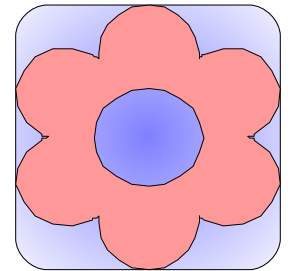
Time taken: 0.08305s [Refresh](#)

- Paris et Berlin demandent à l'Union européenne d'augmenter sa contribution au FMI
- Vol Rio-Paris : la recherche des boîtes noires reprendra à l'automne

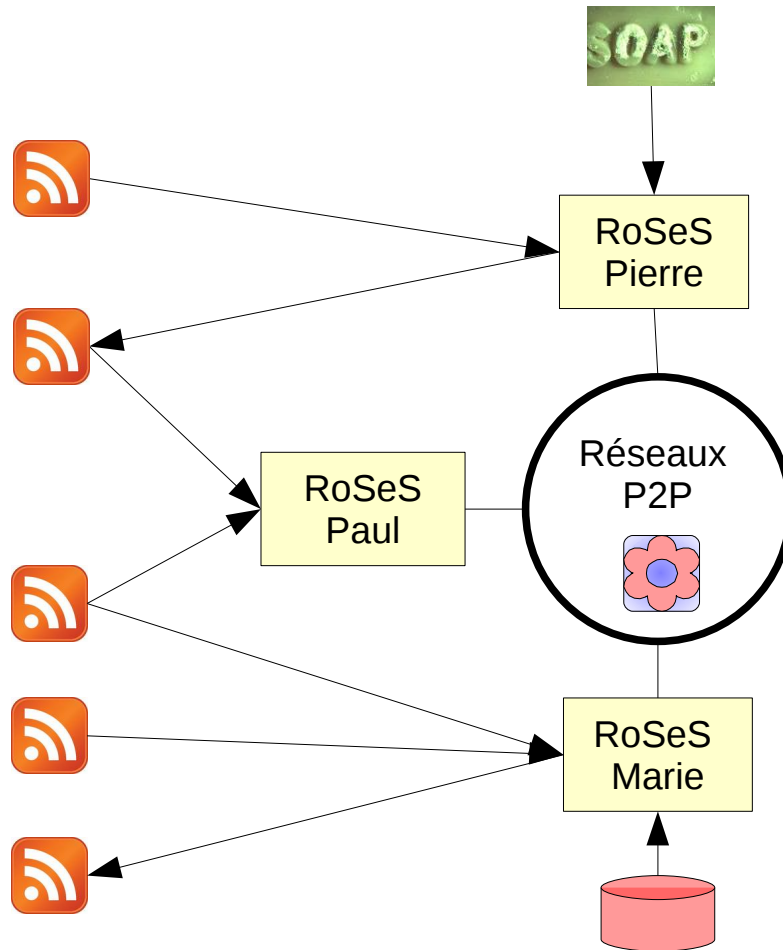
RoSeS : Really Open Simple and Efficient Syndication

Objectif: faciliter la création, l'enrichissement et la personnalisation de flux RSS :

- Modèle de syndication :
 - Flux RSS et données complexes
 - Interrogation et agrégation par requêtes
 - Langages déclaratifs (à la SQL)
 - Exploitation du contexte utilisateur
- Passage à l'échelle :
 - Optimisation bases de données
 - Réseaux Pair-à-Pair



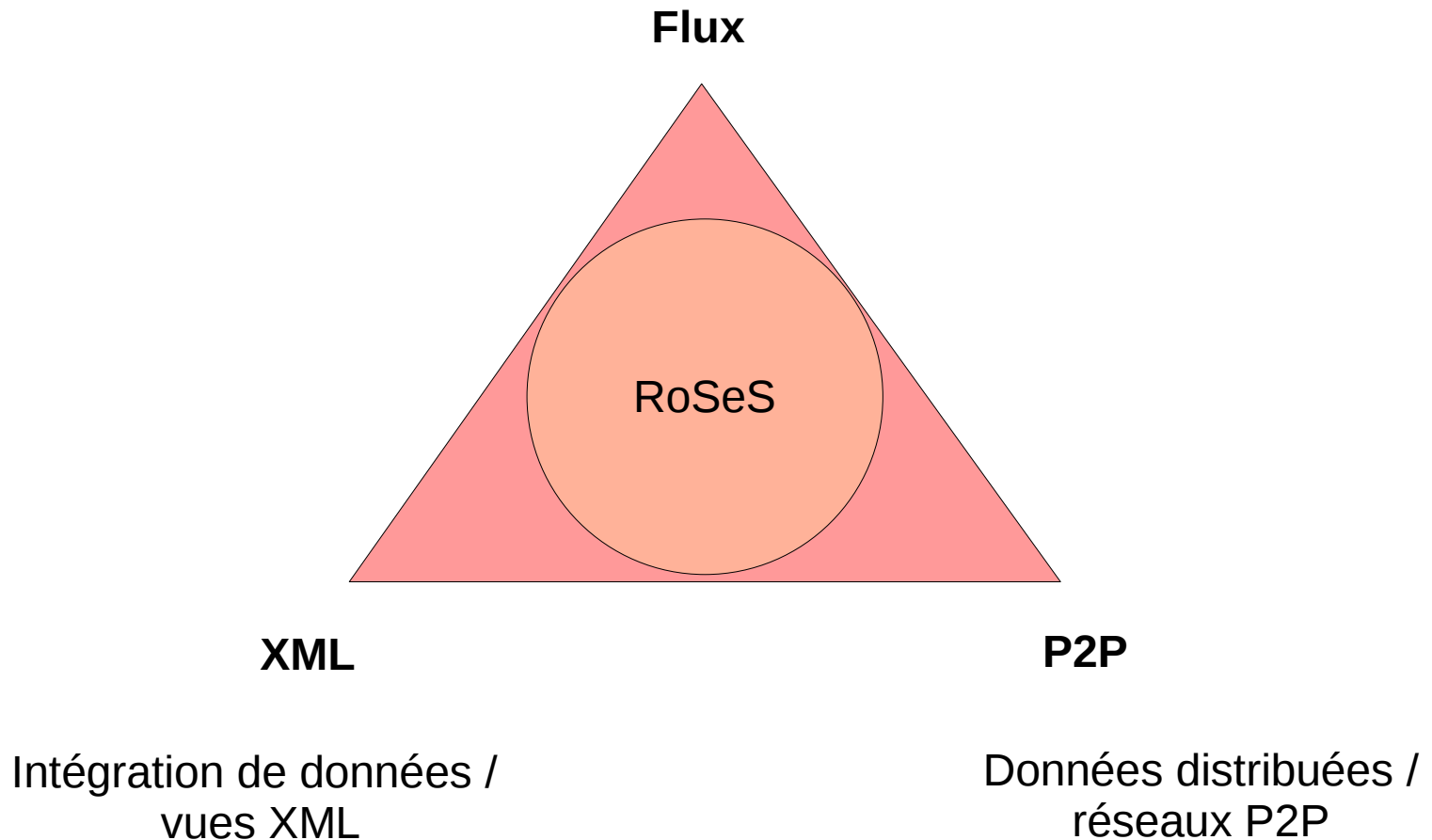
Syndication RoSeS



- Données « mixtes »
 - Flux RSS
 - Bases de données
 - Services web, ...
- Traitements
 - Register
 - Publish
 - Subscribe
- Réseau P2P :
 - Traitements distribuées
 - Optimisation

Contexte scientifique

Flux de données / requêtes continues
séries temporelles



RoSeS

Début – fin :	1 janvier 2008 – 31 décembre 2010
Site Web :	http://www-bd.lip6.fr/roes
Intranet :	http://www-bd.lip6.fr/roesprivate
Gforge :	http://gforge.cnam.fr/gf/project/roes/
Liste de Diffusion :	roes@lip6.fr
Coût complet :	954 k€
Montant alloué :	294 k€
Partenaires :	• LIP6-UPMC (Wisdom) • Cedric-CNAM (Wisdom) • PRISM – UVSQ • LSIS – Univ. de Toulon • 2or3things



RoSes



Activités scientifiques

Workpackages

WP6 : Évaluation et validation

WP5 : Langage RoSeS

WP2 : Modélisation

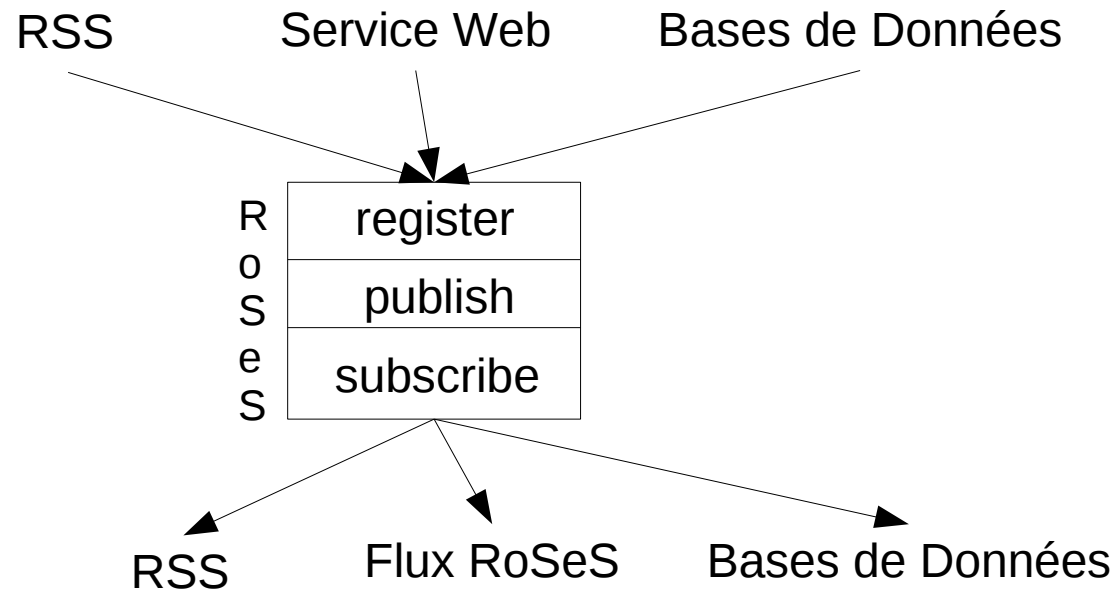
WP1 : Architectures

WP3 : Interrogation

WP4 : Distribution et optimisation

WP5 : Langage RoSeS

- Participants : tous
- Livrables : D5.1
- Résultat :
 - Modèle RoSeS : « RSS » étendu avec contenu XML
 - Langage de syndication



Langage de syndication RoSeS

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

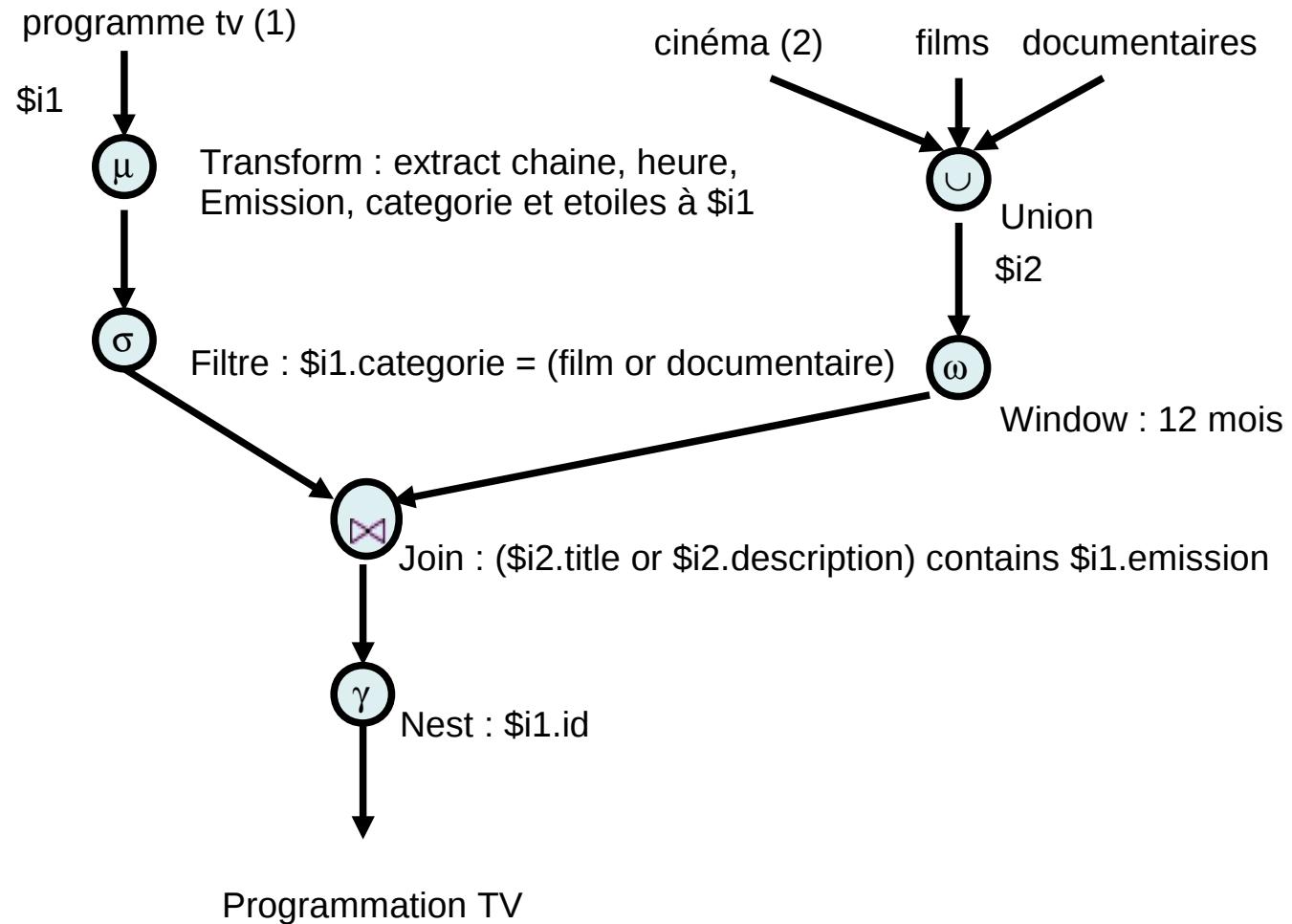
WP2 : Modélisation

- Participants : LIP6, CNAM, LSIS, UVSQ
- Livrables : D2.1, D2.2, D2.3
- Résultats :
 - Modèle et algèbre logique (LIP6, CNAM, LSIS, UVSQ)
 - Classement (ranking) d'items RoSeS (LIP6)

Algèbre RoSeS (D2.2)

- Algèbre RoSeS :
 - D : ensembles d'item RoSeS (RSS étendu)
 - Flux RoSeS : $F \subseteq T \times 2^D$
 - associe à chaque instant $t \in T$ un ensemble d'items $I \subseteq D$
- Requêtes continues :
 - Filtrage, union, transformation
 - Fenêtrage : nombre / temps / contenu
 - Jointure, Agrégation, Group-by
- « Snapshot reducibility » :
 - $op(F) = \{(t, op(I)) \mid (t, I) \in F\}$

Exemple Algèbre



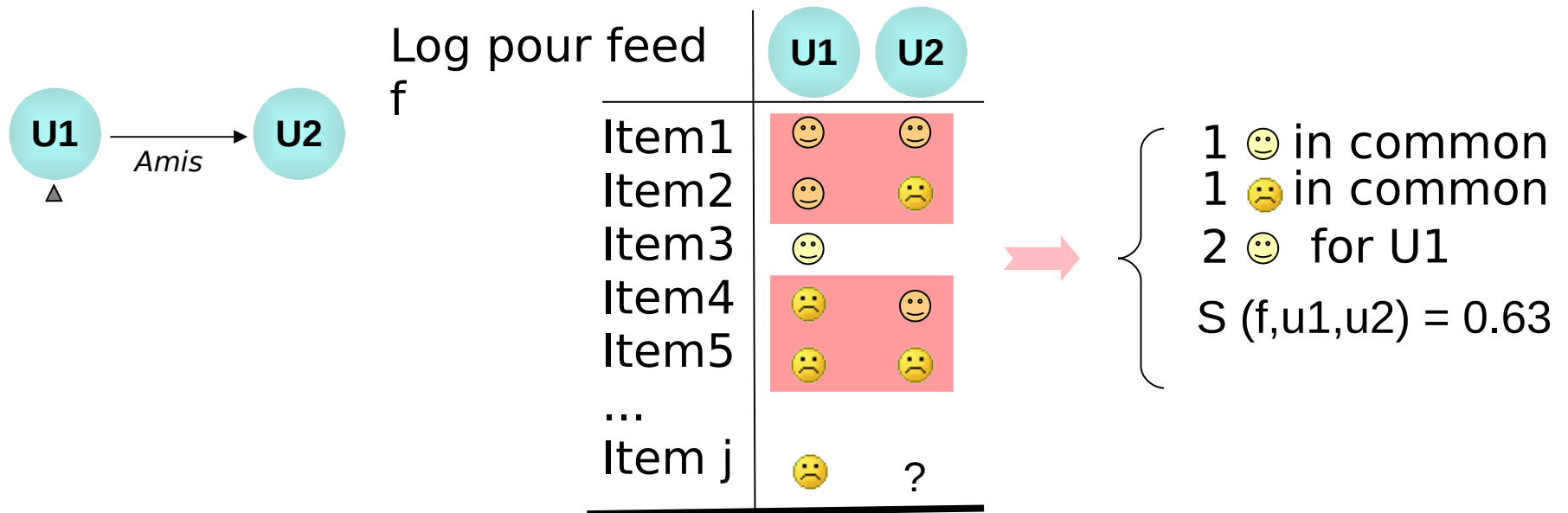
(1) http://www.tele7.fr/tv/flux_rss/cesoir_meschaines

(2) http://www.premiere.fr/recherche/export_rss?type=sortie

Classement d'items

- Objectif : classer les items publiés par un ensemble de flux par la probabilité qu'un utilisateur y soit intéressé
 - Probabilité « historique » d'un item : intérêt moyen aux items du flux d'origine
 - Probabilité « sociale » d'un item : intérêt exprimé par les amis pondéré par leur influence dans le graphe social
- Résultat :
 - Modèle de classement : graphes d'influence
 - Algorithmes pour la propagation des actions de l'utilisateur (click) et le calcul efficace de la probabilité sociale

Classement social

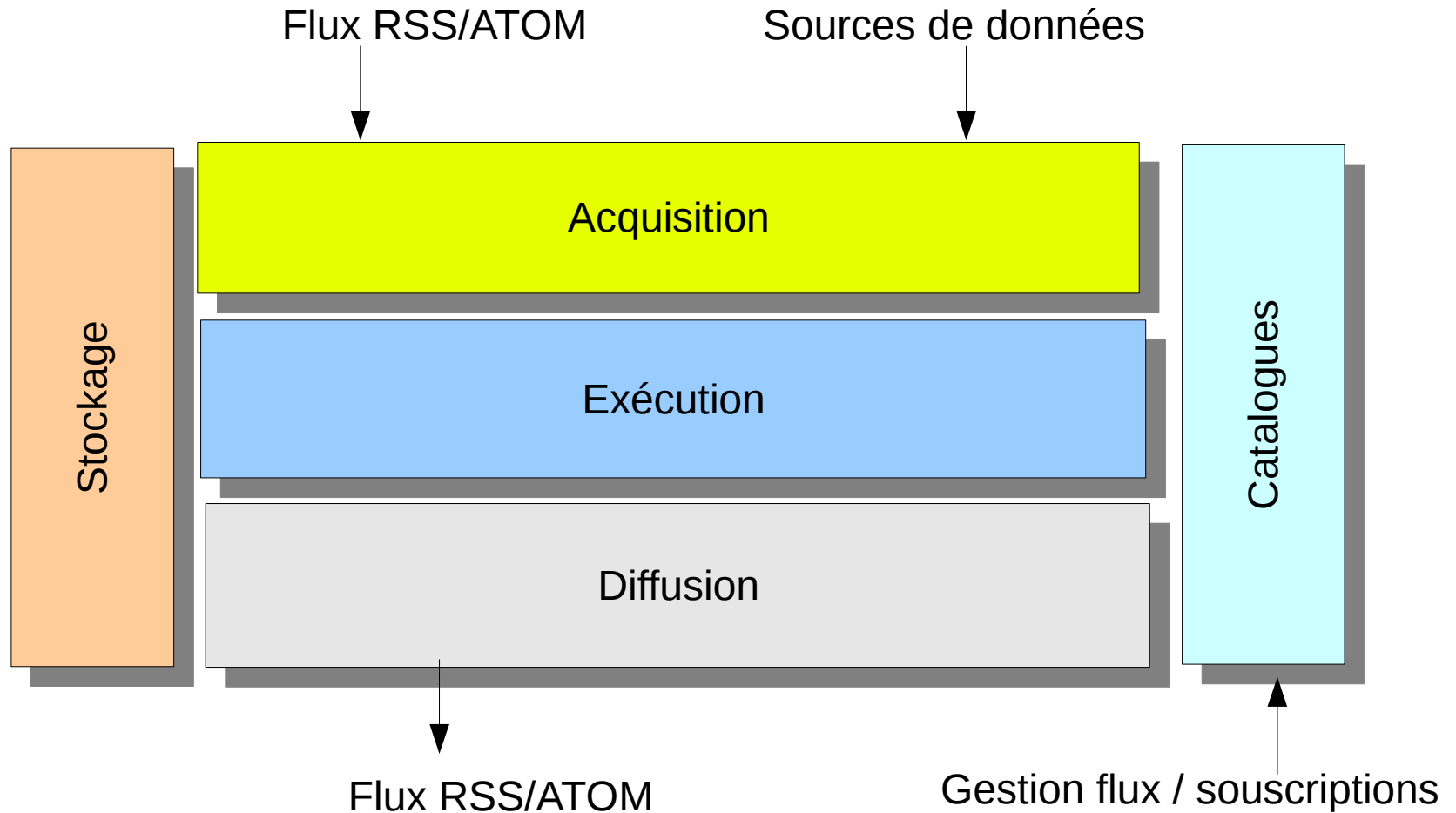


- $S(f, u1, u2)$: combien la présence / l'absence d'intérêt de U1 à l'item j permet de prédire l'intérêt de U2 dans le même item ?
- S : "information mutuelle" (entropie)

WP1 : Architectures

- Participants : tous
- Livrables : D1.1, D1.2, D1.3
- Résultats actuels :
 - spécification fonctionnelle d'un noeud RoSeS (D1.1)
 - deux architectures physiques (D1.2)
 - premières implantations en cours

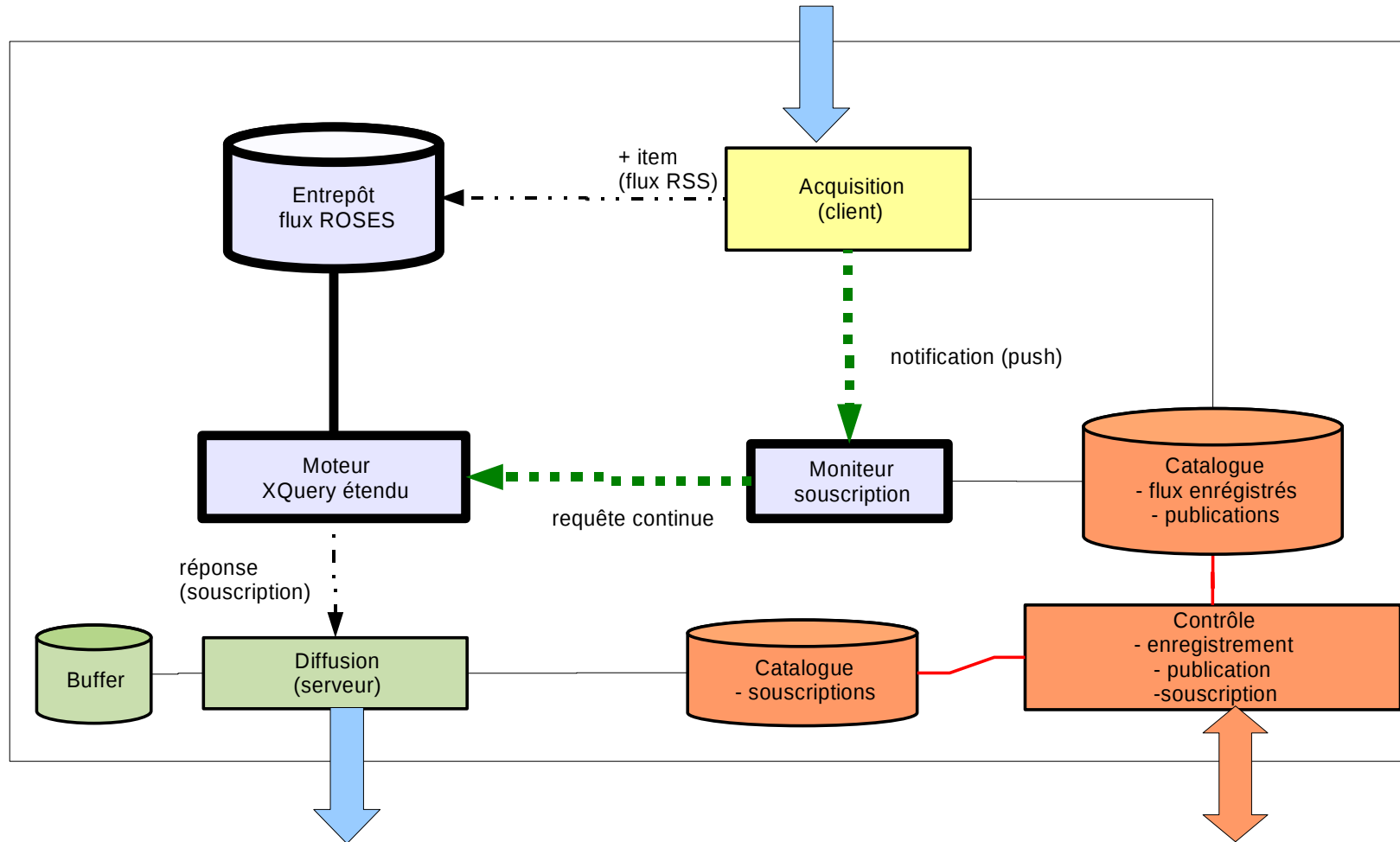
Nœud RoSeS



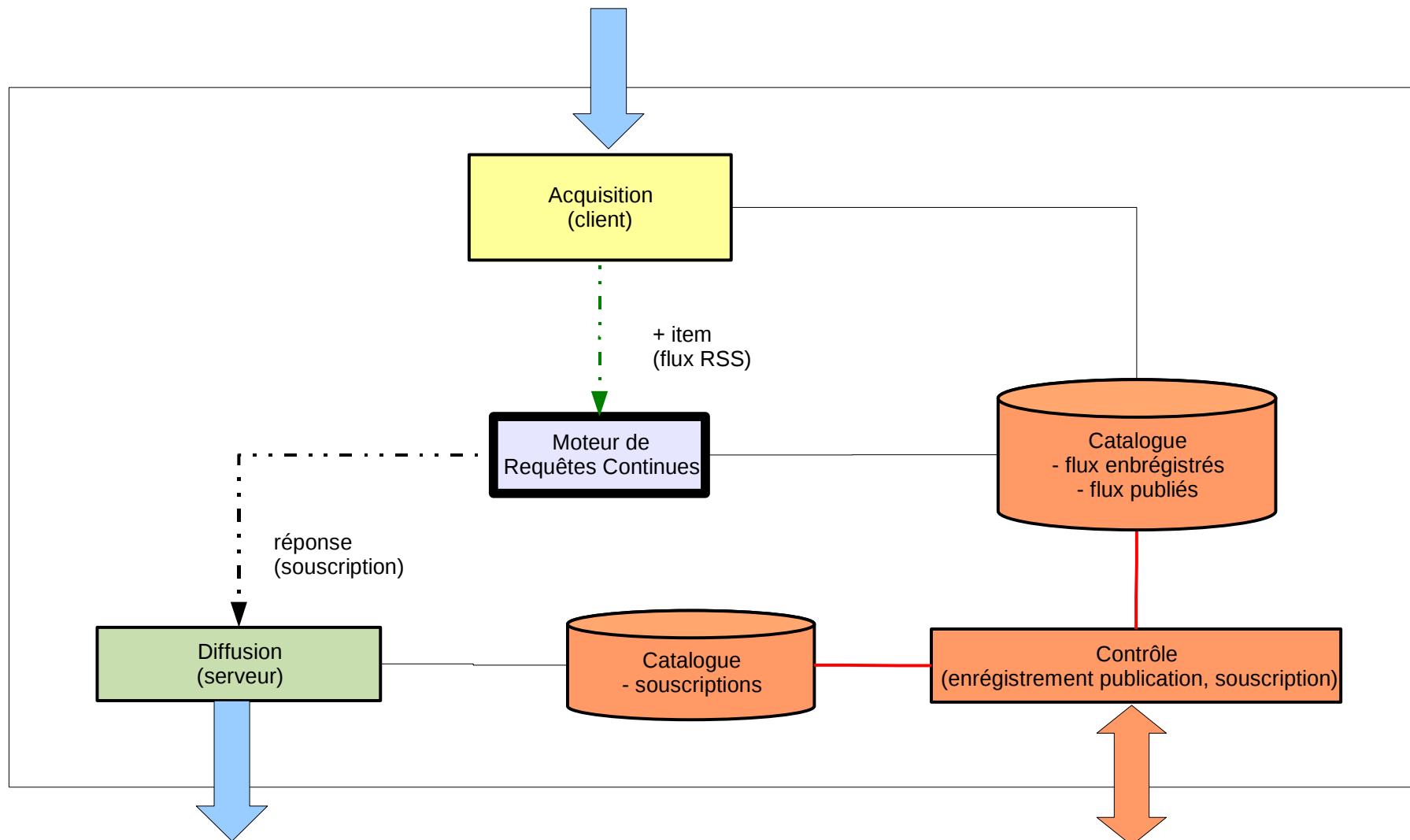
Deux approches de traitement

- Point commun : modèle et algèbre RoSeS
- Approche « entrepôt » (LSIS, UVSQ) :
 - Entrepôt de séquences temporelles alimentées par des flux RSS
 - XQuery étendu avec fenêtrage et agrégation
 - Requête continue : triggers
 - Séries très longues → répartition en P2P
 - Application : analyse boursière
- Approche « requête continue » (LIP6, CNAM):
 - Évaluation en continue (sans stockage)
 - Moteur de requête continue fondé sur des opérations « flux de données »
 - Application : agrégation et enrichissement RSS

Architecture « entrepôt »



Architecture « requête continue »



WP3 : Interrogation

- Participants : tous
- Livrables : D3.1, D3.3, D3.4, D2.2
- Résultats :
 - Module d'acquisition de flux RSS / ATOM avec transformation en flux RoSeS (CNAM, LSIS, 2or3things))
 - Extension du langage XQuery pour l'interrogation de flux RSS (LSIS, PRISM)
 - Modèle et algèbre d'agrégation physique (LIP6, CNAM)

WP3 : Interrogation

Algèbre physique (CNAM, LIP6)

- Flux d'items texte / XML
 - Opérations flux : $\text{flux}^* \rightarrow \text{flux}$
 - filtrage, transformation, union
 - Opérations fenêtrage : $\text{flux} \rightarrow \text{fenêtre}$
 - time-based, count-based, content-based
 - Jointure : $(\text{flux}, \text{fenetre}) \rightarrow \text{flux}$
 - Split : $\text{flux} \rightarrow \text{flux}^*$
- Implantation (en cours) :
 - Files d'attentes et buffers (fenêtres)
 - Multi-thread : thread = ensemble d'opérateurs synchrones

WP3 : Interrogation

Extension Xquery (UVSQ, LSIS)

- Flux RoSeS :
 - séquence temporelle stockée dans un entrepôt XML
- XQuery étendu
 - opérations séquences temporelles : window, agrégation
- Requêtes :
 - statiques : séquence → séquence
 - continues : flux → flux
 - Flux RoSeS : séquence + générateur d'événements
 - Requête :
on <event> in <feed> where <cond_event>
return <XQuery>

WP4 : Distribution et optimisation

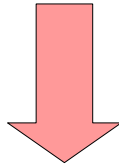
- Participants : LIP6, CNAM, UVSQ
- Livrables : D4.1, D2.4
- Résultats :
 - Modèle de synchronisation adaptative de flux RSS (LIP6)
 - Stratégies d'évaluation de requêtes RSS en P2P (UVSQ)
- Travaux en cours :
 - Optimisation de requêtes continues (CNAM, LIP6)
 - Indexation de souscriptions (CNAM)
 - Optimisation distribuée (UVSQ)
 - Synchronisation distribuée de flux RSS (LIP6)

Synchronisation adaptative

Contraintes :

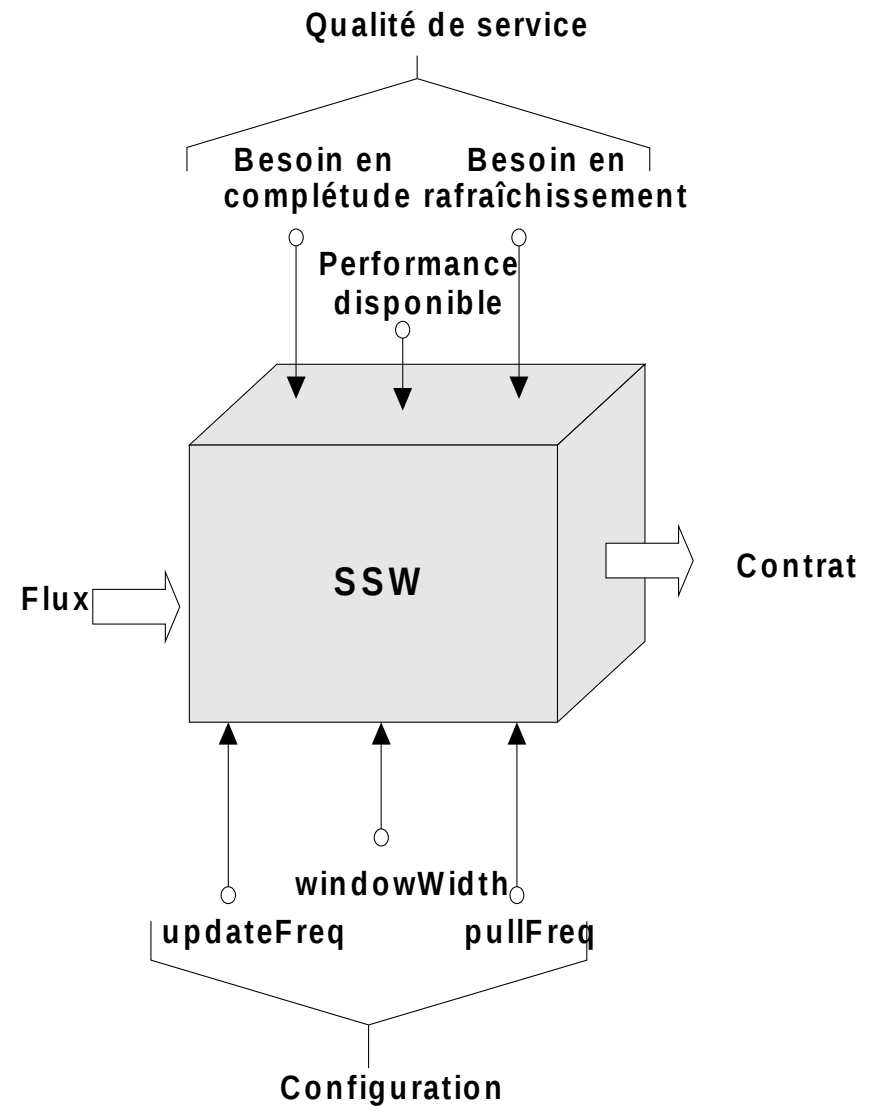
- Besoins de fraîcheur et de complétude
- Fréquence du flux
- Bande passante, mémoire

construire



Contrat synchronisation :

- Fréquence polling
- Taille fenêtre
- Choix protocole (pull / push)



Optimisation publications

- Publications « similaires » (filtrage, fenêtrage)
 - mutualisation, factorisation
 - Multi-opérateurs : $\{op1(F), op2(F)\} \sim op12(F)$
 - Factorisation de sous-expressions (arbres)
- L'ensemble de publications change avec le temps
 - optimisation incrémentale
- Les fréquences des flux interrogés varient avec le temps
 - optimisation dynamique (scheduling)

Optimisation souscriptions

- Problème : passage à l'échelle
 - Des millions de souscriptions (ex. Twitter : 1.9 millions d'abonnés pour CNN)
 - Des dizaines de milliers d'items par jour
 - Une centaine de mots-clés par item
 - 4 mots-clés en moyenne par souscription (source: hitwise pour recherche google)
 - Taille de vocabulaire $|V|=10^3 - 10^4$
- Solution : Arbre de souscription (AKTree)

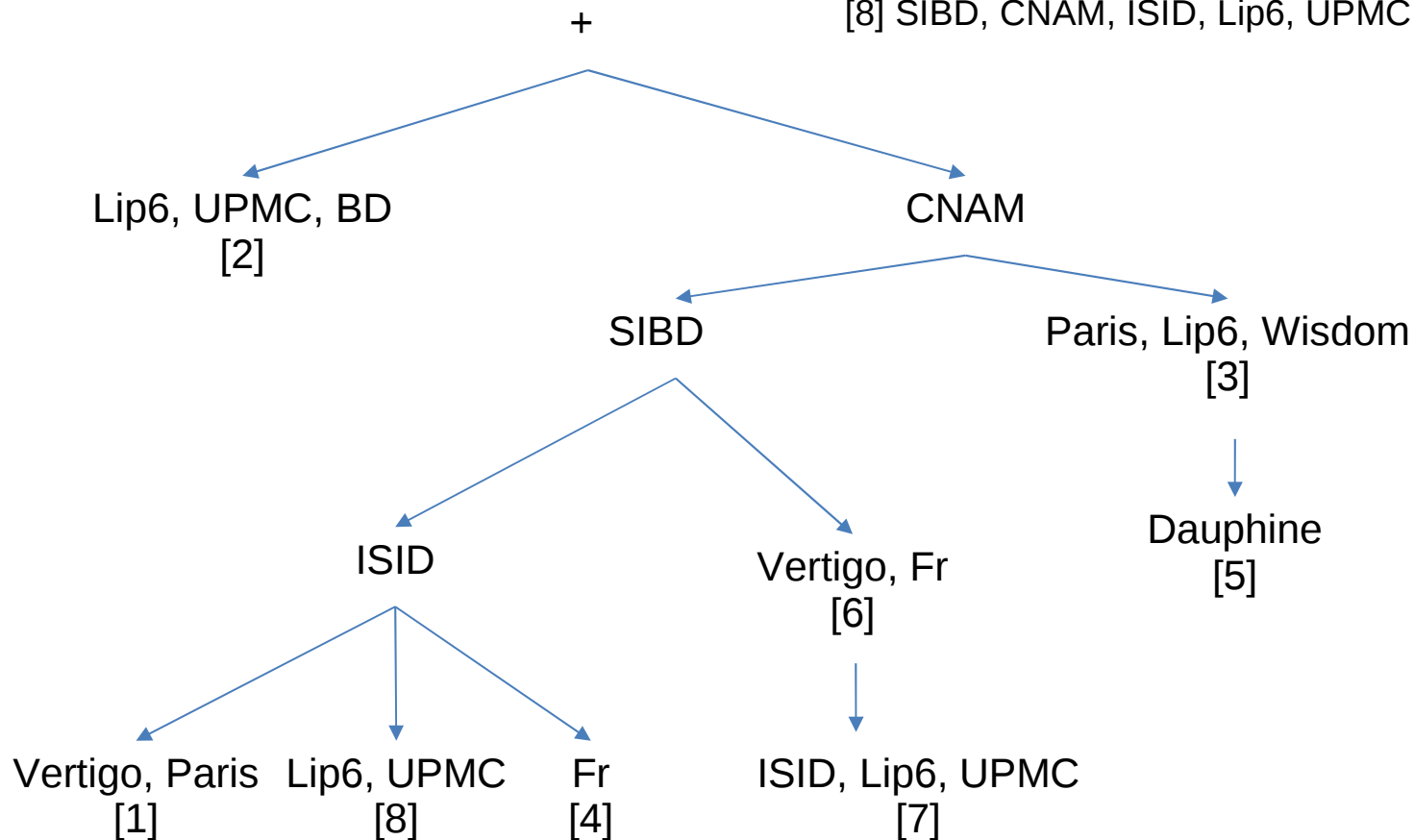
AKTree (CNAM)

- Arbre initial :

- Racine : +
- Un fils / mot clé + souscriptions

- Clustering hiérarchique des frères avec le même ensemble de souscriptions

- [1] SIBD, CNAM, Vertigo, Paris, ISID
- [2] Lip6, UPMC, BD
- [3] CNAM, Paris, Lip6, Wisdom
- [4] SIBD, CNAM, ISID, Fr
- [5] CNAM, Paris, Lip6, Wisdom, Dauphine
- [6] SIBD, CNAM, Vertigo, Fr
- [7] SIBD, CNAM, Vertigo, ISID, Lip6, UPMC, Fr
- [8] SIBD, CNAM, ISID, Lip6, UPMC



Optimisation P2P (UVSQ)

- Réseau de syndication P2P :
 - Indexation de ses ressources (flux, requêtes, caches, etc.)
 - Exécution distribuée de requêtes RSS
 - Délégation transparente et dynamique des souscriptions
- Avantages :
 - Réduction de l'utilisation des ressources globales
 - Équilibrage de charge dynamique
 - Exemple: Adaptation du réseau en fonction des événements ayant une durée de vie limité (ex. Mort Michael Jackson)

WP6 : Évaluation et validation

- Participants : tous
- Livrables : pas de livrables avant T0+36
- Résultats actuels :
 - Ensemble des cas d'usage et exemples de requêtes (Intranet)
 - Simulateurs de flux RSS

WP6 : Évaluation et validation

Travaux effectués :

- Étude de différents scénarios d'applications :
 - Espace d'information personnalisé
 - Partage d'information dans les réseaux sociaux
 - Surveillance de sites web et détection d'événements
- Identification des fonctionnalités principales d'un système d'agrégation de flux RSS
 - Enrichissement
 - Personnalisation
 - Diffusion
 - Archivage
- Préparation d'un benchmark :
 - Exemples de requêtes
 - Simulateurs de flux RSS

Développements

- Simulateur Réseau P2P Roses (synchronisation) : LIP6
- Moteur d'Acquisition RSS/ATOM (extension Blastfeed) : CNAM
- Générateur de flux RSS (benchmark) : LSIS, CNAM
- Moteur de requêtes continues : LIP6
- Indexation de souscriptions : CNAM
- Mini serveur d'un pair (décrit dans le livrable D4.1)
- Réseau P2P indexant des signatures de requêtes XQuery sur des flux : UVSQ
- Fonctions financières sur des TS longues en PHP et Java : UVSQ

Résumé de l'avancement

WP6 : Évaluation et validation

- Simulation
- Applications

WP5 : Langages RoSeS

- Langage de syndication
- Personnalisation et classement

WP2 : Modèle

- Algèbre et sémantique
- Classement (ranking)

WP1 : Architectures RoSeS

- Spécification fonctionnelle
- Architectures physiques

WP3 : Interrogation

- Requêtes continue
- XQuery étendu

WP4 : Distribution et optimisation

- Optimisation requêtes continues
- Indexation requêtes
- Synchronisation
- Réseaux de syndication P2P

Stages

- S. Attrash. Moteur de requêtes continues RSS, stage M2, CNAM/LIP6, 2008
- 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, UVSQ, 2009.
- I. Mami. Optimisation de requêtes continues, stage M2, CNAM/LIP6, 2009
- M. Diouri, Indexation pour la recherche par le contenu textuel de flux, stage M1, CNAM, 2009

Thèses

- Jordi Creus (LIP6-UPMC) : modélisation et traitement continue de flux RoSeS (thèse financée par le projet)
- Bogdan Butnaru (Prism-UVSQ) : P2P et séries temporelles
- Iulian Sandu-Popan (Prism-UVSQ) : Applications de séries temporelles
- Roxana Horincar (LIP6-UPMC) : Synchronisation distribuée de flux RoSeS



RoSes



Fonctionnement du consortium

Organisation

- 4 partenaires universitaires + 1 micro-entreprise
- Réunions de coordination
 - 10 réunions en 2008 / 2009
 - La plupart des réunions réunissaient tous les partenaires
- Archive partagé (Intranet) :
 - Documents scientifiques et administratifs
 - Compte-rendu de réunions
- Serveur Gforge
- Salle de visioconférence (breeze)
- Listes de diffusion

Collaborations

- Première année : étroite collaboration entre tous les partenaires
 - état de l'art, modèle logique, cas d'usage, spécification fonctionnelle, architectures
- Deuxième année : établissement de groupes de travail sur plusieurs thématiques
 - Langage/Modèle Utilisateur, Uses Cases (1,2,3,4)
 - Crawler, Moniteur, Rafraichissement (1,2,4,5)
 - Traitements/Evaluation de Souscriptions (2,3,4)
 - Profils Utilisateurs/Réseaux sociaux (1,2)



ROSeS



Valorisation et Dissemination

Publications et Dissemination

- Publications :
 - G. Gardarin, B. Nguyen, L. Yeh., et.al. "Efficient P2P Processing of Times Series: Application to Stock Investment and Mobile Objects Analysis", BDA'09
 - B. Butnaru, B. Nguyen, G. Gardarin, L. Yeh. « XQ2P: Efficient XQuery P2P Time Series Processing » (démonstration)
- Séminaires LIP6 et Wisdom
- Communication interne

Collaborations extérieures

- Z. Lacroix :
 - Univ. de Arizona et société TGen
 - Préparation d'un contrat NIH intégrant l'approche RoSeS
- V. Christophides
 - Univ. de Crète et FORTH :
 - Modélisation et l'optimisation de graphes de syndication
 - Deux visites d'un mois en 2009 (financement : LIP6 et CNAM)

Bilan

- Problématique riche qui nécessite une expertise dans des domaines nouveaux (flux de données, réseaux sociaux)
- Modèle et architecture RoSeS
- 4 thèses en cours qui traitent des problèmes liés à RoSeS (1seulement est financé par RoSeS)
- Prototypages en cours
- Collaboration étroite sur la modélisation, l'architecture et le développement modules génériques (acquisition, simulateurs)
- Approches particuliers et problèmes spécifiques traitées par les différents partenaires

Objectifs 2009 / 2010

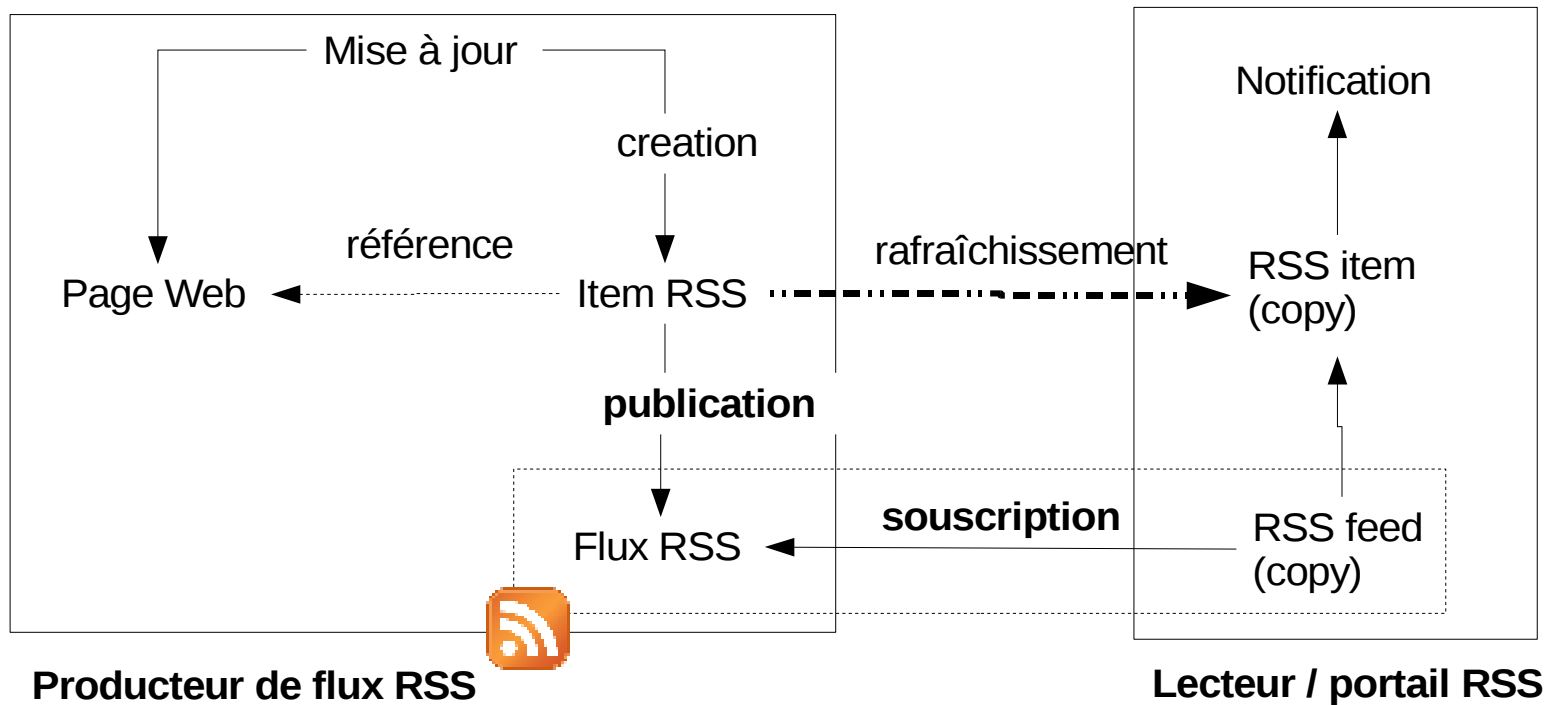
- Publier !
- Prototype(s) intégré(s)
- Journée scientifique / workshop sur les thématiques de RoSeS

Fin

Merci pour votre attention !

ROSeS

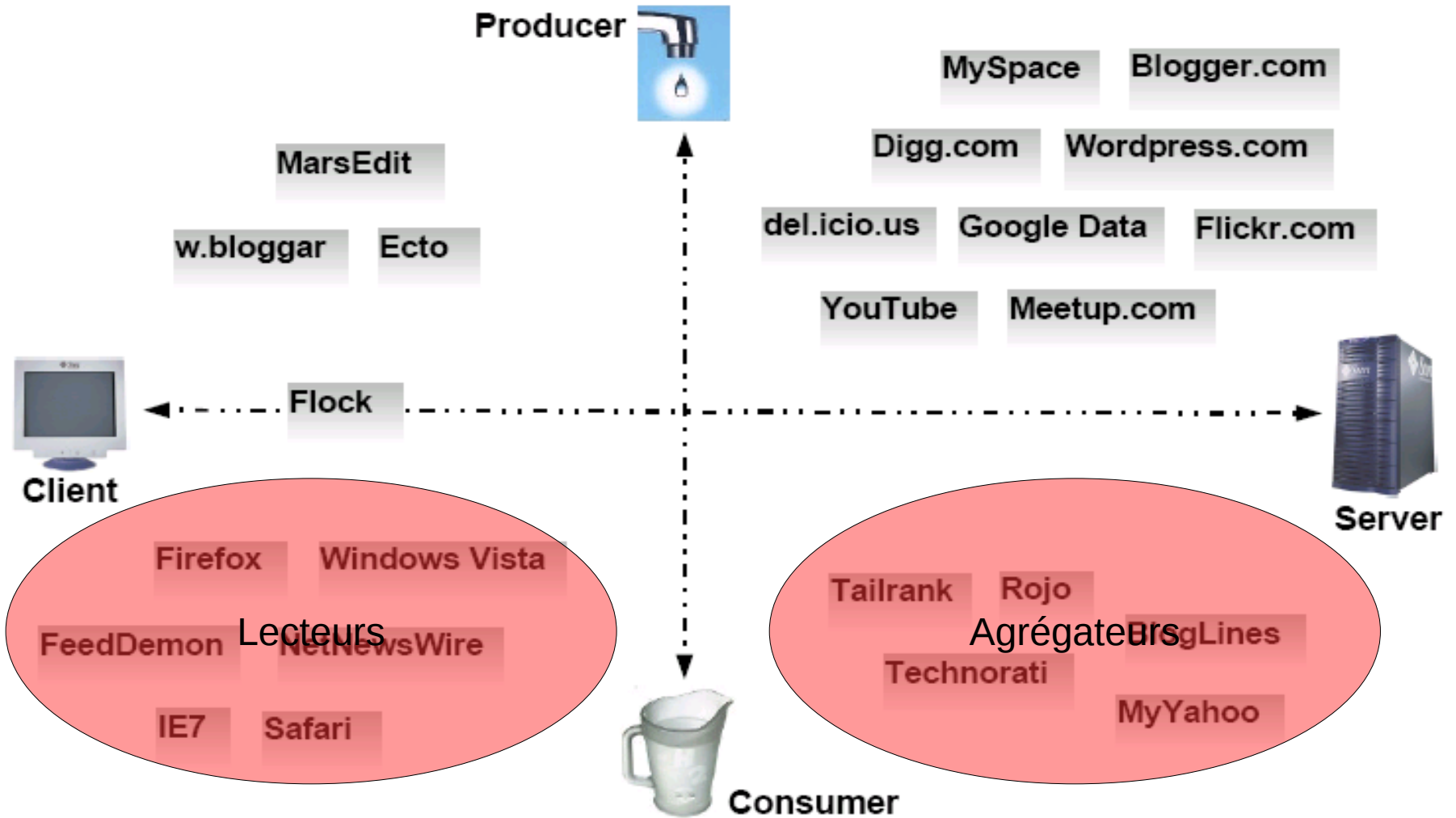
Syndication RSS



Problème de la « Bande Passante RSS »

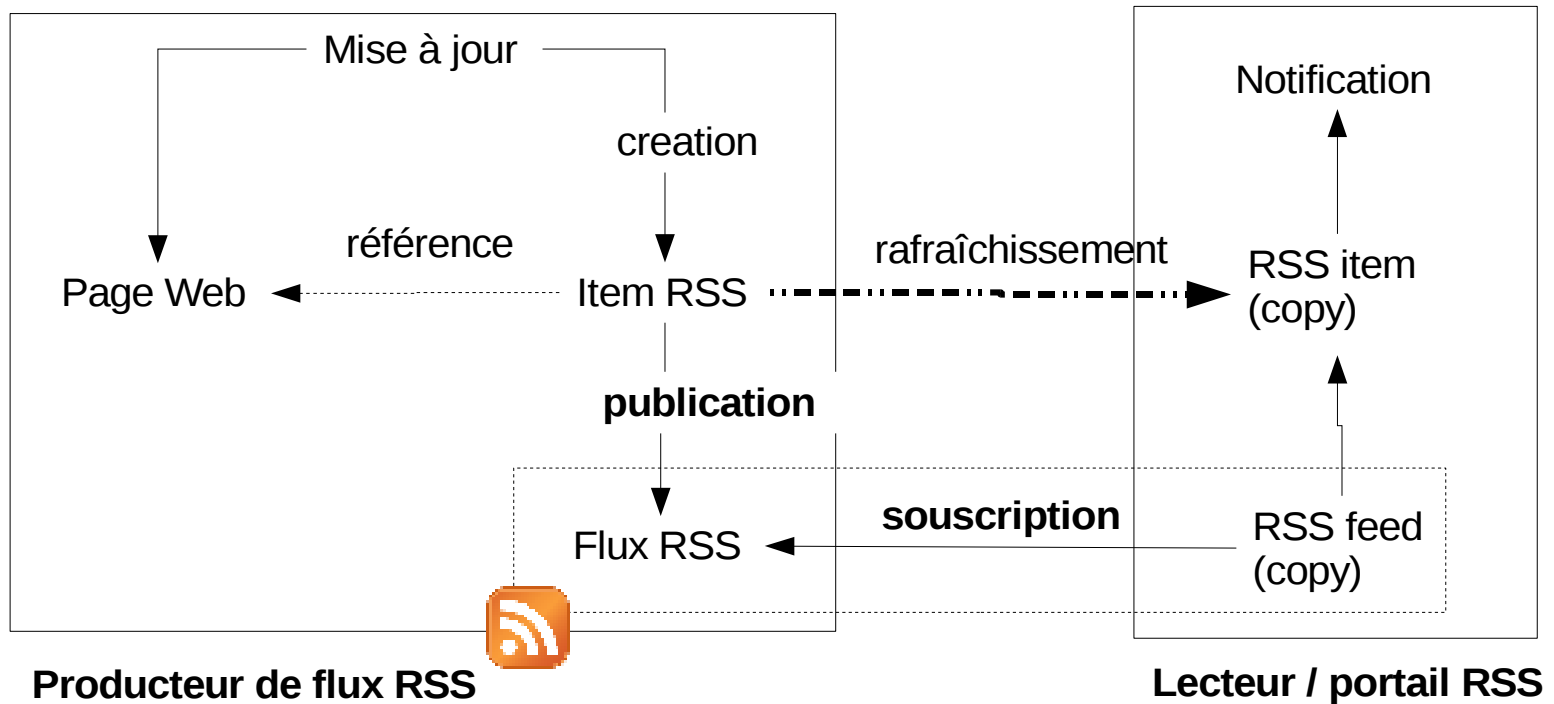
- Problèmes de la bande passante RSS :
 - Fenêtres de taille fixe (fichier) régulièrement et fréquemment mises à jour
 - Plusieurs requêtes HTTP par jour et par utilisateur
 - Échange de données redondant (taille fixe)
 - Trafic continu (24h/24-7J/7) à long terme
- Solutions :
 - Diffusion « lazy » (maintenir l'état de chaque client)
 - Compression de données
 - Estimation fréquence de MAJ (difficile)
 - Architecture trois-tiers centralisées (CAM) et distribuées/P2P (FeedEx, Corona, FeedTree)

RSS Ecosystem



Copyright 2006 David M. Johnson. Released under Creative Commons License Attribution-ShareAlike (<http://creativecommons.org/licenses/by-sa/2.5/>)

Syndication RSS

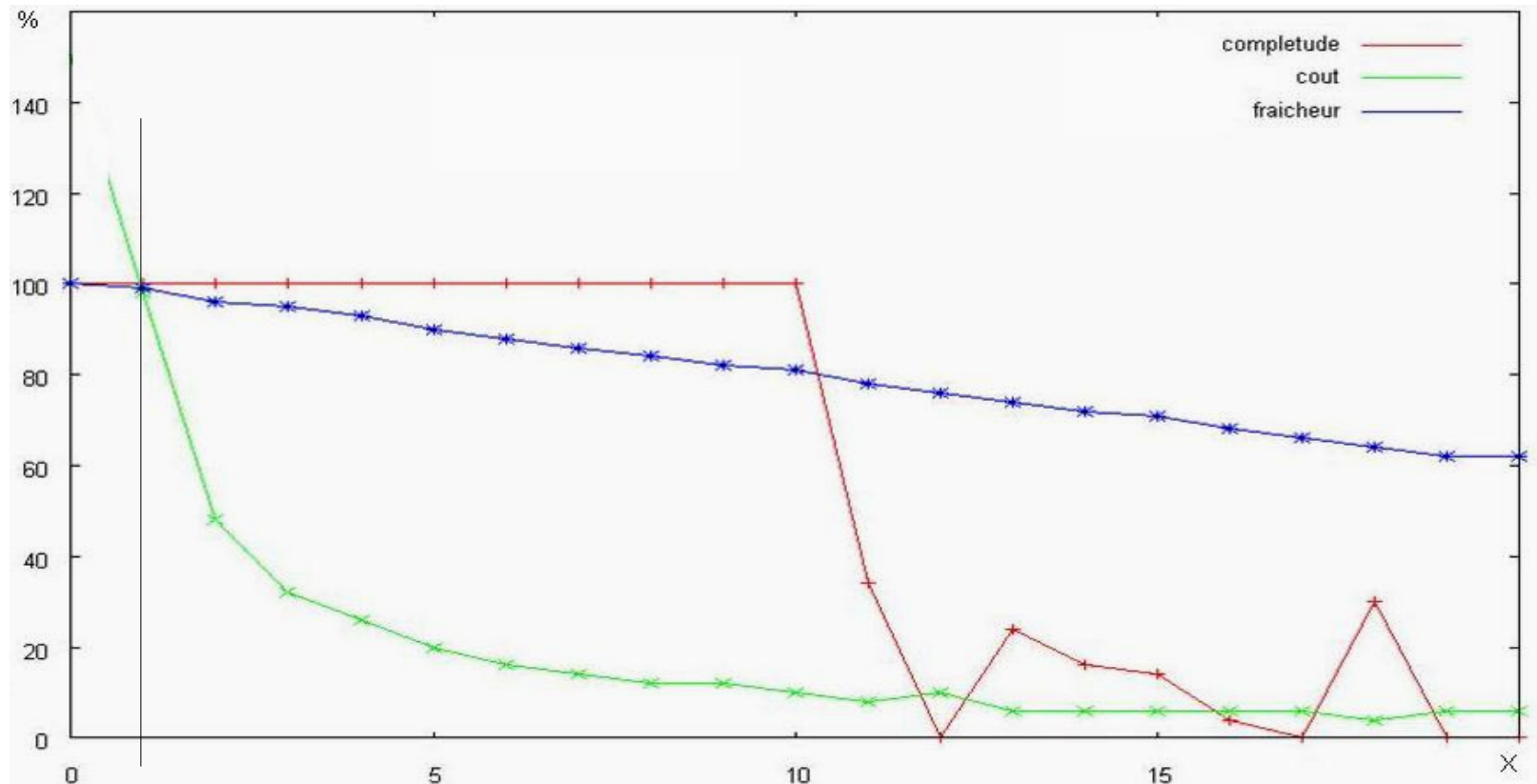


Synchronisation RoSeS

- Contraintes :
 - Qualité : fraîcheur, complétude
 - Matérielles : bande passante, mémoire
 - Infrastructure : client / serveur, réseau
- Optimisations :
 - configuration (statique) :
 - taille fenêtre, fréquence, protocole (pull / push)
 - stratégie (dynamique) :
 - Quoi rafraîchir quand ? (pb. Crawler web)

Expériences

- $T = 10$ = taille de fenêtre de rafraîchissement
- X = # nouveaux items entre deux rafraîchissements
- Coût = $1/X$
- Complétude = $\min \{ 1, T / X \}$
- Fraîcheur = $\max \{ 0, 1 - X / T \}$



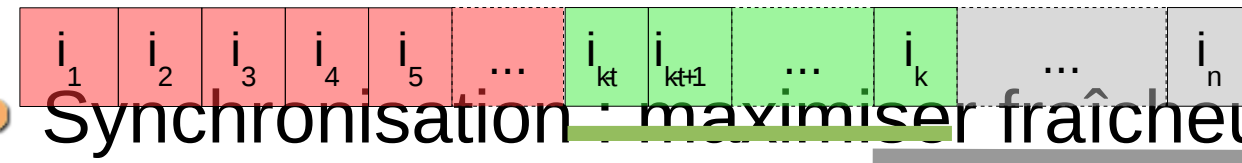
Stratégies de synchronisation

- Réseau de noeuds RoSeS connectés par des liens de souscriptions
- Contraintes / configuration donnée :
 - taille fenêtre, protocole, bande-passante (# rafraichissements / seconde), charge serveur (# connexions / seconde)
- Stratégie de rafraîchissement :
 - identifier à chaque cycle de rafraîchissement les souscriptions à rafraîchir
- Problème :
 - trouver une stratégie qui maximise la fraîcheur des / d'un ensemble de souscriptions

Synchronisation RSS

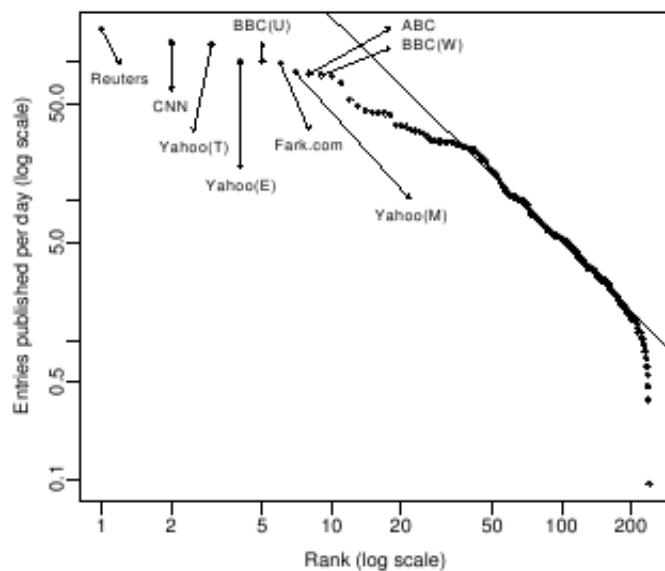
- Fraîcheur : décalage âge client / serveur
- Complétude : % items perdus
- Performance : % items nouveaux par demande de rafraîchissement et # demandes

- Synchronisation : maximiser fraîcheur, complétude, performance

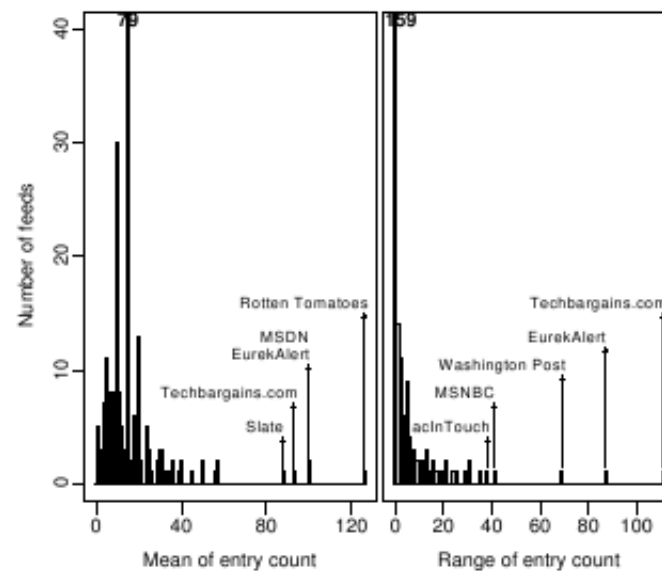


Optimisation publications / souscriptions

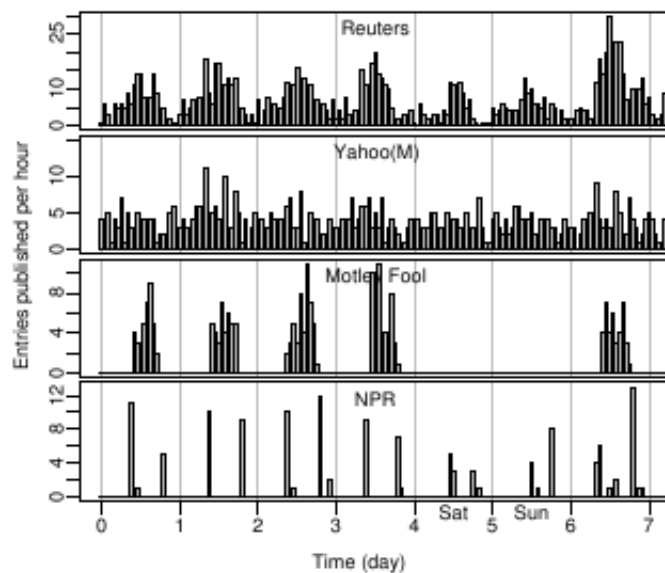
- Publications
 - Définies par les utilisateurs
 - Requêtes continues complexes (agrégation, jointure, union)
 - Nombre limité (dizaines ?)
- Souscriptions
 - Définies par les clients
 - Expressions simples (filtrage)
 - Nombre important (centaines / milliers)



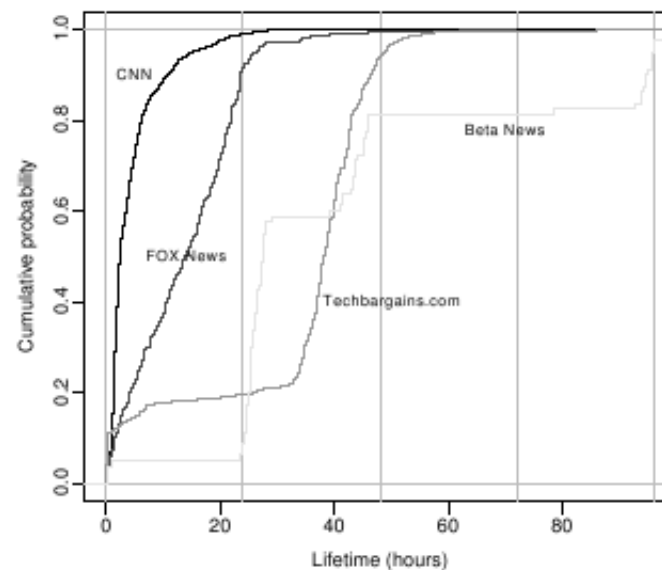
(a) Distribution of publishing rate. The line is fitted between ranks 30 and 200.



(b) Histograms about entry count. Two long bars go up to 79 and 159, respectively.



(c) Change of publishing rate. Each bar is one hour wide. Notice different y-axis scales.



(d) Cumulative distribution of entry lifetime. Vertical guidelines indicate days.

Figure 2: Statistics of news feed publishing

Source : Feedex : collaborative exchange for news feeds - Seung Jun, Mustaque Ahamad. WWW06