

Gestion d'information sur le Web

Serge Abiteboul

INRIA Saclay
LSV, ENS Cachan & LRI, Paris 11

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE



centre de recherche
SACLAY - ÎLE-DE-FRANCE

Webdam

Information et informatique

L'informatique est la science de l'information et du traitement de l'information

L'informatique - contraction d'information et automatique - est le domaine d'activité scientifique, technique et industriel en rapport avec le traitement automatique de l'information par des machines telles que les ordinateurs, les consoles de jeux, les robots, etc. [Wikipedia]

Le sujet ici

- La gestion d'information
- Et plus particulièrement, de l'information sur le Web

S. Abiteboul – INRIA Saclay – Mars 2010



INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE



centre de recherche
SACLAY - ÎLE-DE-FRANCE

Organisation

Le Web grand public

Les standards de la gestion de données sur le Web

Tendances

Parallélisme: grappes de machines

Vérification d'applications et workflows

Conclusion

S. Abiteboul – INRIA Saclay – Mars 2010



INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE



centre de recherche
SACLAY - ÎLE-DE-FRANCE

Le Web grand public

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE



centre de recherche
SACLAY - ÎLE-DE-FRANCE

Internet et le Web

Internet permet de transférer des flux d'information entre machines connectées au réseau (TCP)

- C'est le point de départ de tout, e.g., la messagerie électronique

1^{er} Web (Tim Berners-Lee – HTML)

- On donne des noms à des documents hypertextes
- On les publie & on peut les lire et avec les liens passer de l'un à l'autre
- Moteur de recherche: trouver les documents « les plus pertinents » contenant des mots clés

1^{er} Web: Information textuelle universelle pour des humains

Le Web: Echange d'information entre humains & machines

S. Abiteboul – INRIA Saclay – Mars 2010



INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE



centre de recherche
SACLAY - ÎLE-DE-FRANCE

Vertige des tailles des données et du Web

Accélération de la taille

BD classiques: gigabyte (10^9 bytes)

Maintenant : tera (10^{12}), peta (10^{15})

Mondiales 2010 (?) : zeta (10^{21})

Le Web actuel

Milliards de pages

Millions de serveurs

Milliards d'objets communicants
(e.g., smartphones)

Ça va très vite

- Années 70: Arpanet
- Années 80: Web, email
- Années 90: Google, Wiki
- Années 00: Facebook, P2P, Blogs, Web 2.0, Twitter
- Années 10: ???

S. Abiteboul – INRIA Saclay – Mars 2010



INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE



centre de recherche
SACLAY - ÎLE-DE-FRANCE

Success stories sur le Web

Google: gestion des pages du Web

Amazon, eBay: catalogues de vente sur le Web

Facebook: informations personnelles et communautés

Emule: musique en ligne

Flickr: base de données de photos

Myspace: pages Web

YouTube: vidéos

Wikipedia: dictionnaire

Meetic: fiches individuelles

Dailymotion: vidéos

Twitter: communication

Quel est leur point commun ?

Gestion d'information sur le Web

S. Abiteboul – INRIA Saclay – Mars 2010



La gestion d'information

Un grand succès de l'informatique du 20ème siècle

- Industrie et recherche académique
- Le modèle relationnel
- Des tableaux à deux dimensions sur des **serveurs centralisés**

Technologie des systèmes de gestion de bases de données

Logique → Algèbre → Optimisation

Nom	Père	Sexe
Adam		M
Eve		F
Caïn	Adam	M
Abel	Adam	M
...		

S. Abiteboul – INRIA Saclay – Mars 2010



INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE



centre de recherche
SACLAY - ÎLE-DE-FRANCE

Logique et algèbre

Médiateur intelligent entre utilisateur & machine

Logique

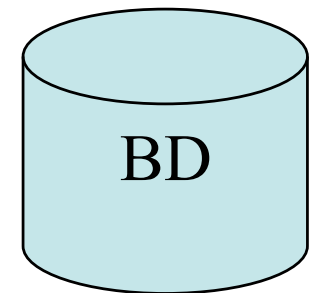
- On exprime une requête (ses besoins) sous forme déclarative - calcul de prédicats du 1^{er} ordre
- Regain d'intérêt pour la logique des modèles finis
- Logique aussi pour les contraintes d'intégrité et d'autres trucs

Algèbre relationnelle

- On traduit la requête en une expression avec des opérateurs de l'algèbre (projection, jointure, sélection...)

On optimise cette expression et on l'évalue

- Algorithmes sophistiqués



S. Abiteboul – INRIA Saclay – Mars 2010



Pourquoi ça marche?

Ces requêtes sont évaluable de manière très efficace

Complexité classique

- En temps exponentiel dans la taille de la requête
- Seulement polynomial dans la taille des données

Complexité parallèle

- AC0, circuits de profondeur constante & taille polynomial dans les données
- Cela prouve que plus on a de processeurs, plus on peut aller vite

Et on fait mieux en utilisant des structures d'accès

- En faisant des calculs « à l'avance »
- Hachage, Arbres B

S. Abiteboul – INRIA Saclay – Mars 2010



Comment on y est arrivé?

Une nouvelle fonctionnalité ou l'amélioration d'une fonction existante

Des modèles mathématiques simples

Des algorithmes intelligents pour optimiser l'usage des ressources
(processeurs, *mémoire*, *disque*, réseau)

Un engineering solide

Et tout cela en tenant compte des progrès des machines

S. Abiteboul – INRIA Saclay – Mars 2010

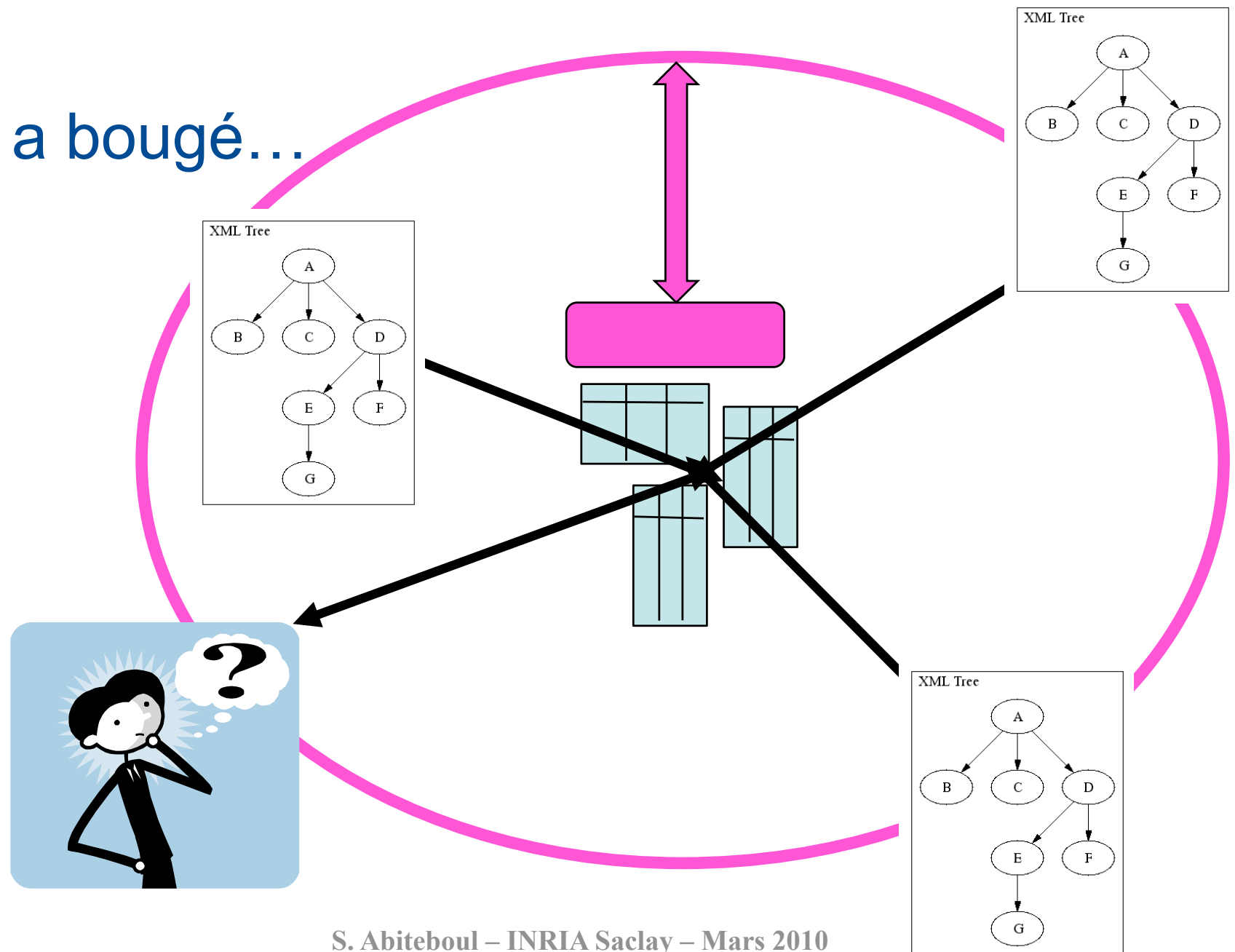


INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE



centre de recherche
SACLAY - ÎLE-DE-FRANCE

Ca a bougé...



S. Abiteboul – INRIA Saclay – Mars 2010

Ca a bougé

Les données sont hétérogènes (format, structure, métadonnées, ontologies, multimédia, etc.)

Les serveurs de données sont hétérogènes, distribués, autonomes, parfois mobiles (BDs, PDA, téléphones, objets communicants, senseurs...)

L'information est changeante, imprécise, incohérente parfois

Parfois grande échelle: Millions de serveurs, terra octets de données, milliards d'objets communicants

S. Abiteboul – INRIA Saclay – Mars 2010



INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE



centre de recherche
SACLAY - ÎLE-DE-FRANCE

Pour quelles sources d'information?

Base de données sur un serveur

Une machine outil

Un système de fichier

Un équipement télécom

Un serveur Web

Un jouet

Un PC

Etc.

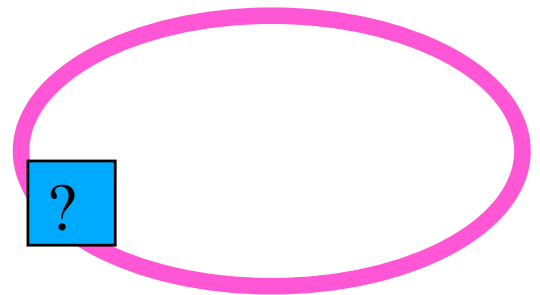
Un PDA

Un smartphone

Un senseur

Un objet communicant -
domotique

Une voiture

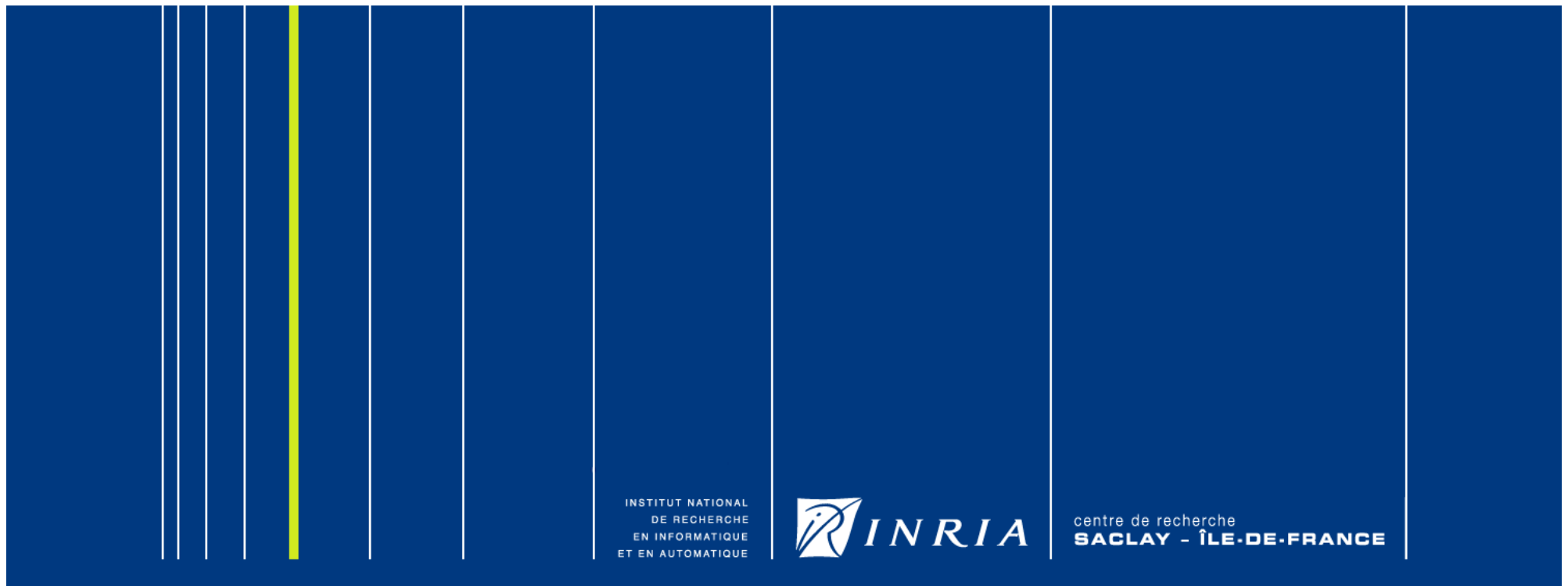


N'importe quel objet ou logiciel
connecté au réseau avec de
l'information à partager

S. Abiteboul – INRIA Saclay – Mars 2010



Un accès universel à l'information ...



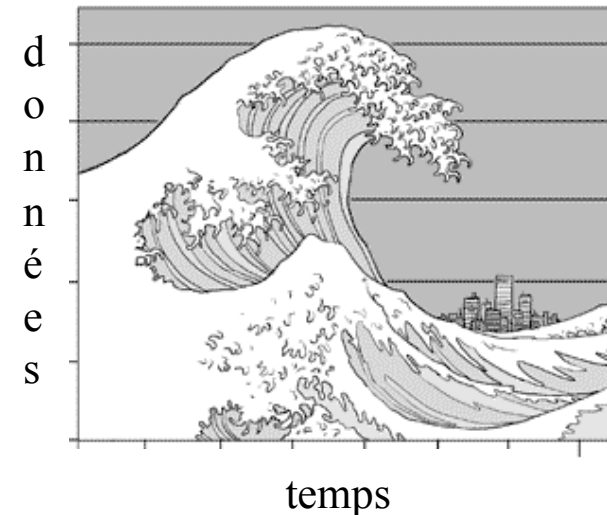
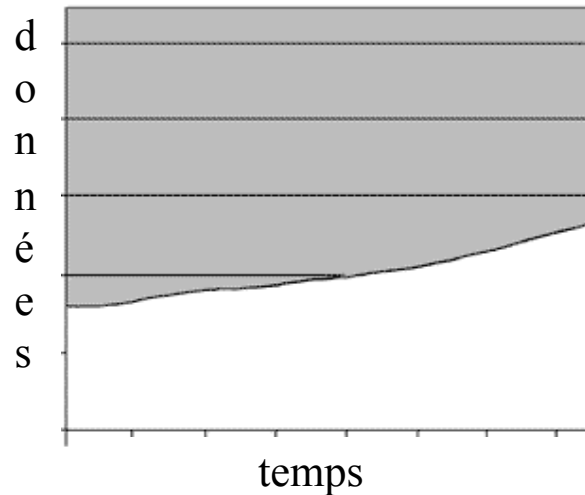
Le risque: se noyer sous un océan de données

De plus en plus facile de publier

De plus en plus de données publiées

De plus en plus difficile de trouver l'information

De plus en plus difficile de l'avoir à temps



Des risques sociétaux – un autre sujet

Internet et le Web offrent d'énormes possibilités mais avec les énormes risques d'une révolution de cet ampleur

La fracture numérique

Mise en danger de l'espace privé – Big brother, droit à l'oubli...

Utilisation du Web par

- Les terroristes, les pédophiles, les criminels en général

Le risque de devenir idiot en laissant le Web choisir pour nous

Le risque de perdre contact avec la vraie vie (vs. les mondes virtuels), les vrais amis (vs. les amis FaceBook)...

D'autres...

S. Abiteboul – INRIA Saclay – Mars 2010



INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE



centre de recherche
SACLAY - ÎLE-DE-FRANCE

Les standards de la gestion de données sur le Web

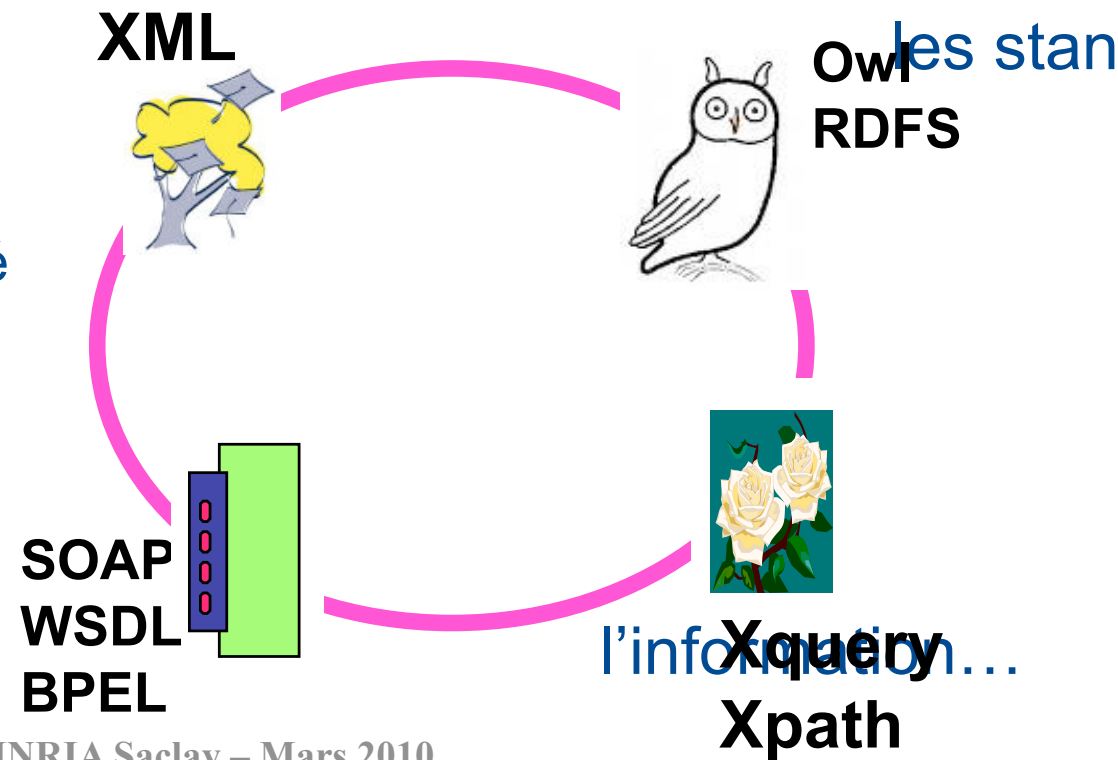
C'est d'abord une histoire de standards

L'information résidait sur des îles avec des formats, des langages de programmation, des applications, des systèmes d'exploitations différents

Mais ça a changé avec

- XML : données
- Xquery : requêtes
- SOAP : calcul distribué
Services Web - BPEL
- Owl : sémantique

Accès uniforme et universel à



S. Abiteboul – INRIA Saclay – Mars 2010



Les standards rapidement

XML : des arbres ordonnés et étiquetés

Xquery et Xpath : « logique » pour poser des requêtes

- Inspiré de SQL/OQL (standards pour données relationnelles et objet)
- Mélange le contenu et la structure: BD et recherche d'information

Services Web : des fonctions

- Possibilité d'activer des **fonctions** sur des serveurs Web distants
- Une famille de standards: SOAP, WSDL, UDDI
- Workflows et chorégraphie de services: BPEL

RDF : logique pour décrire des connaissances

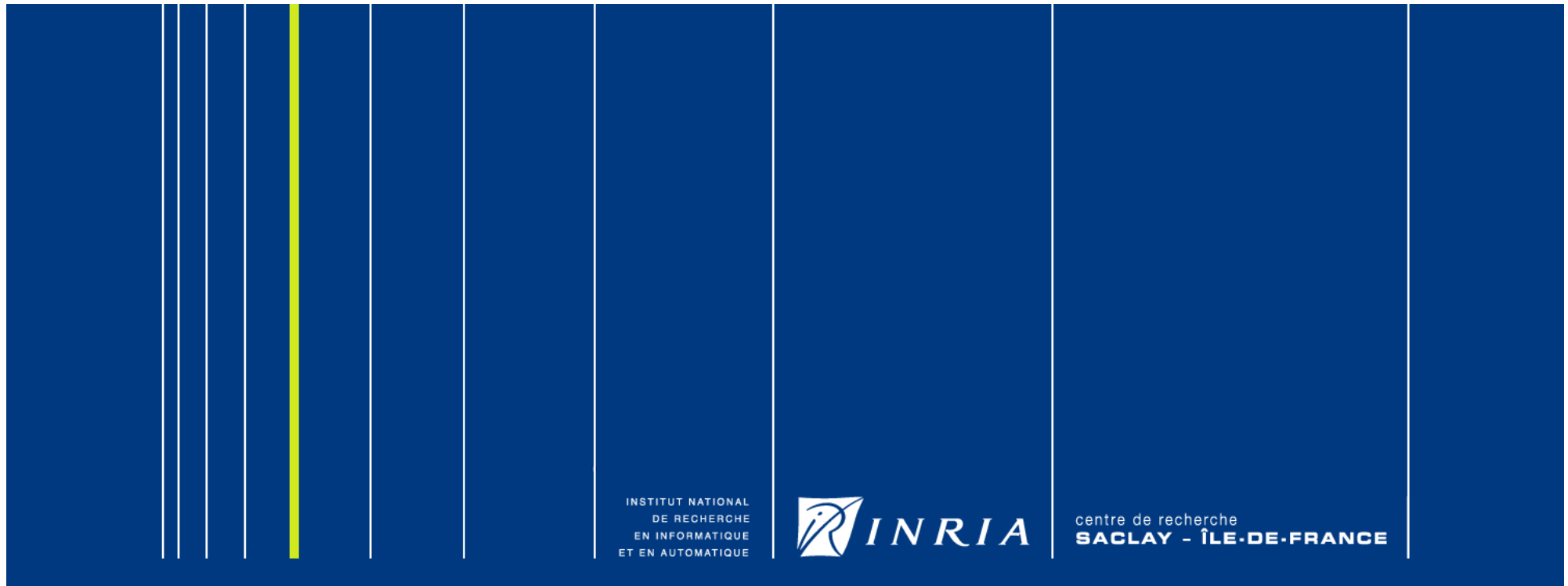
- Pour décrire la sémantique des données – représentation

Ces standards permettent de mettre en commun de l'information

S. Abiteboul – INRIA Saclay – Mars 2010



Tendances



Tendance: Web sémantique

Publier des connaissances au lieu de publier du texte

Un programme peut utiliser ces connaissances

- Aider l'utilisateur
- On peut poser des questions précises et obtenir des réponses précises

Problèmes durs

- Obtenir ces connaissances
- Extraction: linguistique et multilinguisme

S. Abiteboul – INRIA Saclay – Mars 2010



INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE



centre de recherche
SACLAY - ÎLE-DE-FRANCE

Tendance – Web 2.0 (un buzzword)

Le Web n'est pas juste fait pour obtenir des données

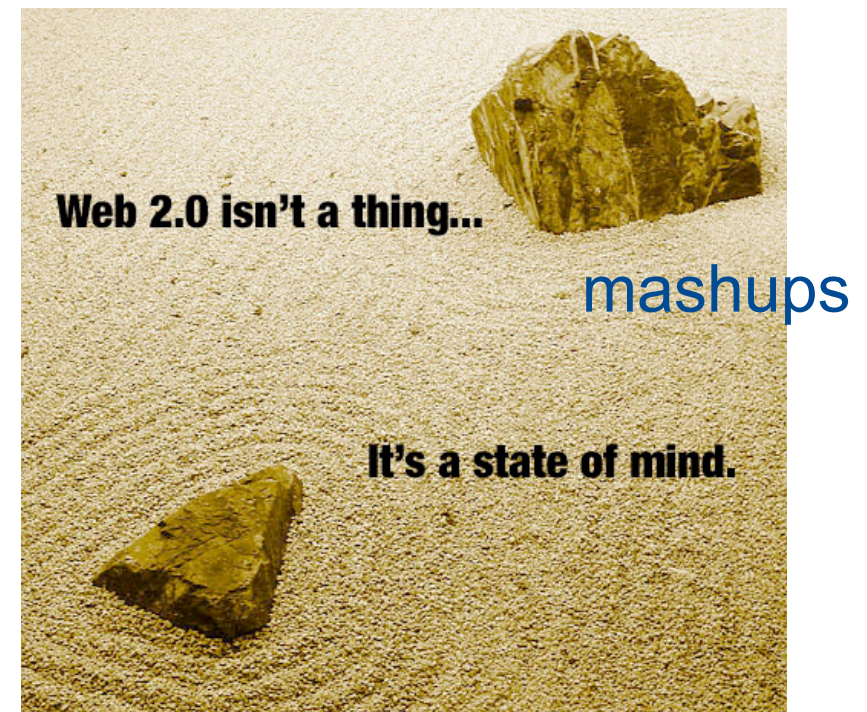
Mot clé: interaction, communauté

Pub: Wiki, Wikipedia, Twitter

Développement d'applications:

Multimédia

- musique, photos, vidéos...



S. Abiteboul – INRIA Saclay – Mars 2010



INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE



centre de recherche
SACLAY - ÎLE-DE-FRANCE

Tendance – Web 2.0 – les réseaux sociaux

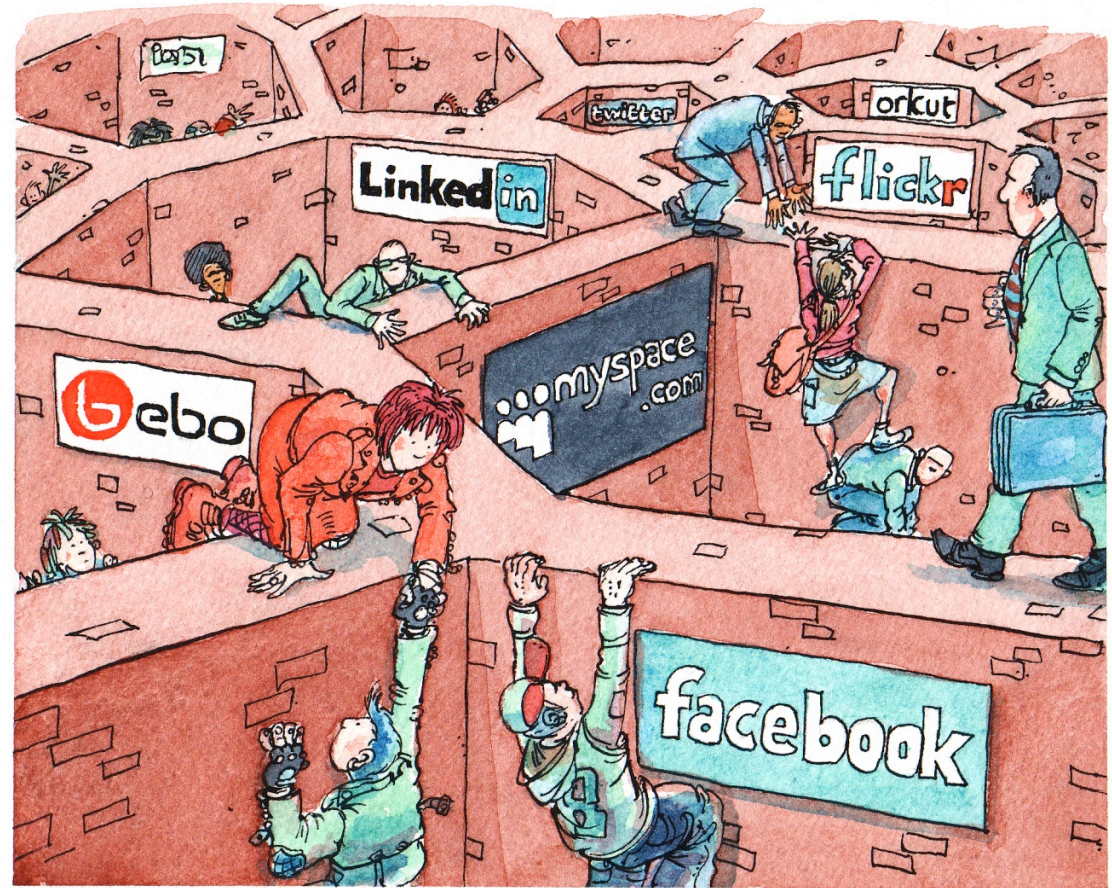
Partage & échange d'information

Les jeunes se sont appropriés le Web avec Facebook

Les entreprises...

Les associations...

Les séniors...



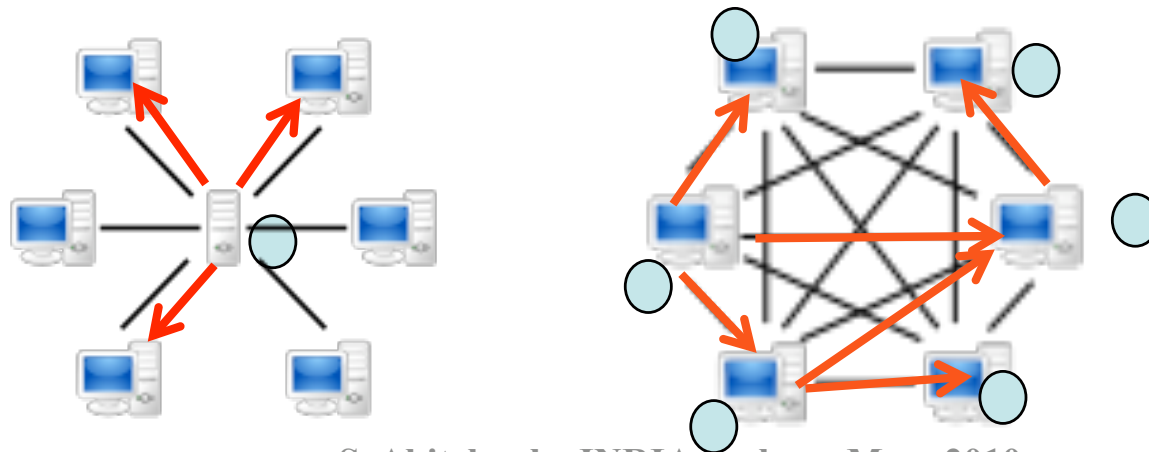
S. A.

Tendance – pair-à-pair

Utiliser les ressources du « réseau » sans passer forcément par des serveurs

- Communication: Skype
- Calcul: seti@home
- Stockage: emule

Puissance du parallélisme: disponibilité, performance, etc.



S. Abiteboul – INRIA Saclay – Mars 2010



Tendance – et d'autres

Web des smart phones & mobilité

- Iphone

Web des objets et intelligence ambiante

- De plus en plus d'objets communicants
- Objets avec des étiquettes genre RFID

Web du 3D et des mondes virtuels

- Second Life

S. Abiteboul – INRIA Saclay – Mars 2010



INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE



centre de recherche
SACLAY - ÎLE-DE-FRANCE

2 zooms

1 sur le parallélisme

1 sur la vérification d'applications

Plein d'autres sujets importants pour le Web qu'on ne traitera pas

- Contrôle d'accès et mises-à-jour
- Surveillance du web: voir webcontent.fr
- Fouille de données, e.g., recommandation

S. Abiteboul – INRIA Saclay – Mars 2010

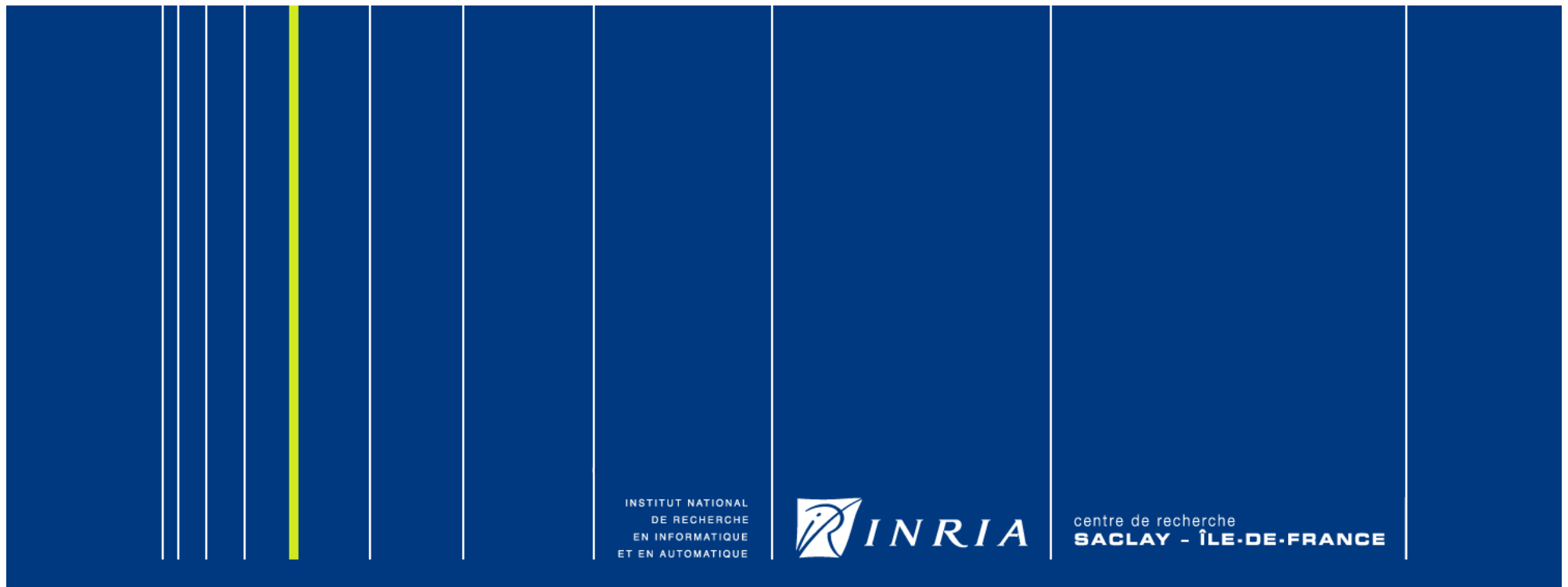


INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE



centre de recherche
SACLAY - ÎLE-DE-FRANCE

Parallélisme: grappes de machines



Exemple: Indexeur du Web

Gérer un index qui pour chaque mot donne la liste des pages qui le contiennent – problème d'échelle

- Des dizaines de milliards de requêtes de recherche par mois
- Des milliards de pages
- Index est du même ordre de grandeur que les pages indexées

Choisir quelles pages montrer d'abord – Pagerank

- Des produits de matrices gigantesques

Solution Google: parallélisme et mémoire

- Des fermes avec des milliers de machines
- Mettre l'index en mémoire (la mémoire est moins chère)

S. Abiteboul – INRIA Saclay – Mars 2010



Digression – Comment on y est arrivé

Une nouvelle fonctionnalité ou l'amélioration d'une fonction existante

Des modèles mathématiques simples

Des algorithmes intelligents pour optimiser l'usage des ressources
(processeurs, **mémoire**, disque, **réseau**)

Un engineering solide

Et tout cela en tenant compte des progrès sur les machines

Même chose pour le modèle relationnel

S. Abiteboul – INRIA Saclay – Mars 2010



INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE



centre de recherche
SACLAY - ÎLE-DE-FRANCE

Les fermes de machines

Performance à cause du parallélisme

Disponibilité

Passage à l'échelle

- On peut s'adapter aux besoin
 - Plus de clients/requêtes: ajouter des machines,
 - Plus de pages indexées: ajouter de la mémoire

Coût bien moindre qu'avec des gros calculateurs

Difficulté: administration de la ferme

S. Abiteboul – INRIA Saclay – Mars 2010



INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE



centre de recherche
SACLAY - ÎLE-DE-FRANCE

Informatique dans les nuages

Utiliser des ressources informatiques disponibles sur le Web (par exemple via services Web)

- Quelles ressources: disque, calcul, applications, etc.
- Fournisseur: Google, Yahoo!, Amazon, Salesforce...

L'utilisateur n'a pas à gérer les systèmes/applications

Complicé techniquement mais c'est le problème du fournisseur de service

Exemple de problème technique: comment gérer des milliers de BD sur des centaines de machines

S. Abiteboul – INRIA Saclay – Mars 2010



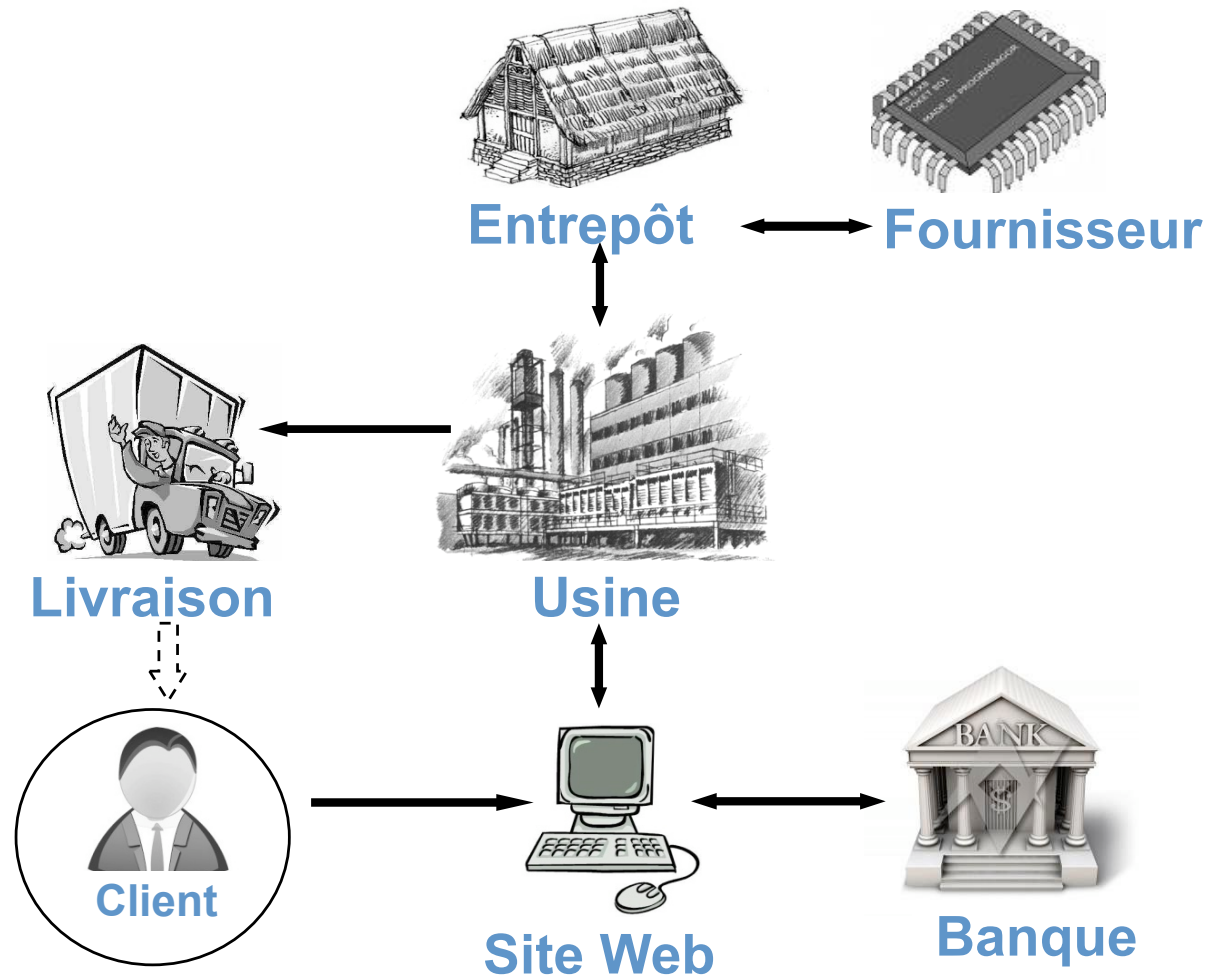
INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE



centre de recherche
SACLAY - ÎLE-DE-FRANCE

Vérification d'applications & workflow

Exemple: Chaine de fabrication de Dell



Problème

Problème général du Web: qualité de service souvent inacceptable en particulier pour l'industrie

- Fiabilité, contrôle d'accès, sécurité...

Complexité de cette application particulière

- Grand nombre de machines autonomes
- Très dynamique (mises-à-jour, échanges de données, communication)
- **Difficile à concevoir & vérifier & surveiller**
- **Difficile de garantir la qualité de service**

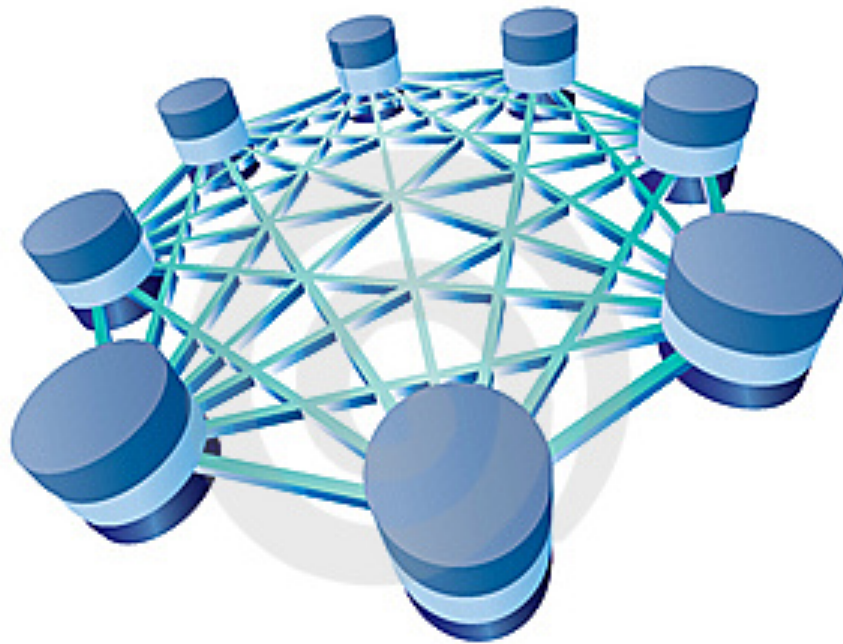
Faire collaborer deux types de systèmes

- | | |
|--------------------------------------|--------------------------|
| • Gestion de données | Système de gestion de BD |
| • Gestion du séquençement des tâches | Système de workflow |

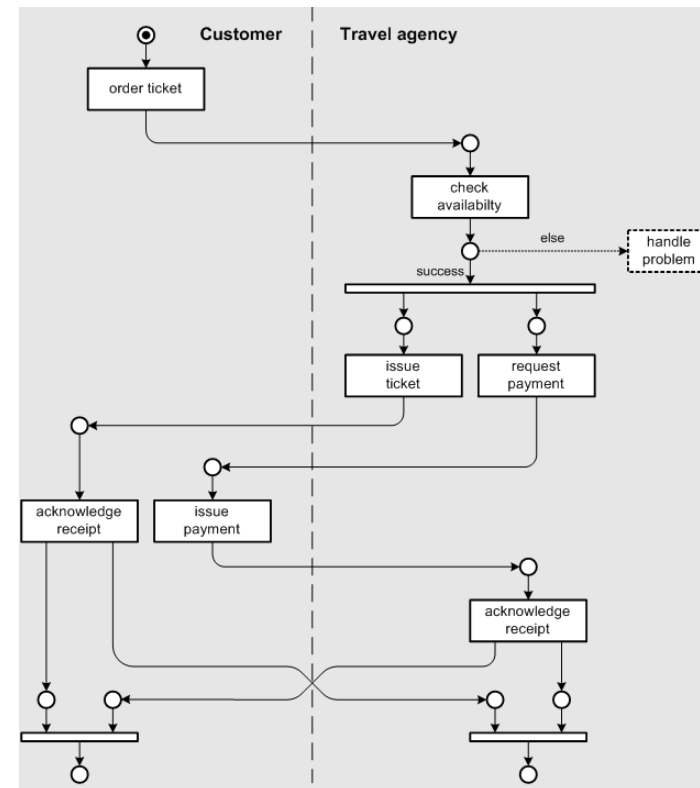
S. Abiteboul – INRIA Saclay – Mars 2010



Un réseau de systèmes de bases de données



Des systèmes de workflow



Une des difficultés : les faire collaborer

S. Abiteboul – INRIA Saclay – Mars 2010



Idée simple

Tout baser sur l'échange de documents entre les différentes entités

Concept introduit par IBM

Notion d'*artifact*

- Un artifact est un document qui représente l'état d'une activité
- Il peut se déplacer dans l'entreprise
- Son évolution est contrôlée par des contraintes

Exemples

- Une commande, un voyage, un projet, une réservation, un employé, une réunion, un dossier...

S. Abiteboul – INRIA Saclay – Mars 2010



INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE



centre de recherche
SACLAY - ÎLE-DE-FRANCE

Une commande

Dans le site Web

webOrder id=7787780
 Customer
 Name: John Doe
 Address: Sèvres
 Order selection: **on-going**
 Ref: PC 456
 Factory: *undecided*
 Parts: *not-active*
 orderDate: 2009/07/24
 Site: http://d555.com
 Payment: *pending*
 Delivery: *not-active*

Dans l'usine

webOrder id=7787780
 Customer
 Name: John Doe
 Address: Sèvres
 Order selection : *committed*
 Ref: PC 456
 Factory: Milano
 Parts: **on-going**
 orderDate: 2009/07/24
 Site: http:// d555.com
 Payment: *done*
 Bank-account ...
 Delivery: *not-active*

Chez le livreur

webOrder id=7787780
 Customer
 Name: John Doe
 Address: Sèvres
 Order selection : *committed*
 Ref: PC 456
 Factory: Milano
 Parts: *done*
 orderDate: 2009/07/24
 Site: http:// d555.com
 Payment: *done*
 Bank-account: CEIF-4457889
 Delivery: **on-going**
 Address: Orsay

S. Abiteboul – INRIA Saclay – Mars 2010

Artifacts – Différents points de vue

Méthodologie pour concevoir l'application

Logiciel pour faciliter les développements

Compilateur transformant des spécifications en un code

Besoin semblables: workflows scientifiques

- Séquencement de tâches
- Gestion de la provenance des données utilisées

S. Abiteboul – INRIA Saclay – Mars 2010



INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE



centre de recherche
SACLAY - ÎLE-DE-FRANCE

Conclusion

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE



centre de recherche
SACLAY - ÎLE-DE-FRANCE

Ma recherche personnelle

Gestion d'information

- Trop complexe pour bricoler
- Fondements mathématiques solides dans le cadre relationnel: des tableaux à deux dimensions sur des serveurs centralisés

Gestion d'information sur le Web

- Encore plus complexe
- Besoin de fondements mathématiques sur des données plus complexes distribuées
- Logique, théorie de la complexité, théorie des automates...
- Vérification d'applications de gestion de données distribuées

S. Abiteboul – INRIA Saclay – Mars 2010



INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE



centre de recherche
SACLAY - ÎLE-DE-FRANCE

Prévisions

Sur le Web, soyons modestes !

Ça va ralentir – on a fait le plus facile & on arrive aux sujets durs

- Linguistique, gestion de connaissances
- Vérification et qualité de service

Ça va continuer à bouger – créativité humaine

- Web sémantique : Web de connaissances plutôt que de texte
- Web 2.0 : un Web plus interactif, plus communautaire
- Web des objets, du pair-a-pair, des téléphones intelligents...

Rappel : pour le Web, on a toujours sous-estimé la créativité humaine

S. Abiteboul – INRIA Saclay – Mars 2010



INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE



centre de recherche
SACLAY - ÎLE-DE-FRANCE



	INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE	 INRIA	 centre de recherche SACLAY - ÎLE-DE-FRANCE	
--	---	--	---	--