Toutes les données du monde en un clic

Ioana Manolescu-Goujot
INRIA Saclay, équipe LEO

7 octobre 2011

SOMMAIRE

1

Les bases de données

On dit que tout a commencé avec les banques

Comptes bancaires: bien avant les ordinateurs!

Voici vos comptes sur l'année:

On dit que tout a commencé avec les banques

La bourse aussi!

Vue de l'extérieur:

On dit que tout a commencé avec les banques

La bourse aussi!

Vue de intérieur:

Que demande-t-on de la gestion d'un compte courant?

1. Quand on fait un achat,
soit je suis débité **et** le vendeur est crédité,
soit rien ne bouge.

Atomicité

2. Quand j'achète, cela ne peut que faire diminuer le montant sur mon compte.
Je n'ai qu'une date de naissance et une adresse principale.
Je n'ai pas moins de 18 ans et pas plus de 120…

Consistence

3. Si je fais des achats en même temps que mon mari et qu'un remboursement INRIA arrive en même temps, cela ne va pas produire des erreurs

Isolation

4. La banque ne va pas oublier combien d'argent j'ai
Ni combien je leur dois…

Durabilité

Que demande-t-on de la gestion d'un compte courant?

1. Quand on fait un achat,

ACIDité: propriétés des transactions "normales"

Théorie des transactions: Jim Gray '70

Prix Turing 1998 pour "contributions fondatrices aux bases de données et à la gestion des transactions"

Atomicité

Consistence

uer le montant sur mon compte.

e principale.

…

Isolation

on mari et qu'un remboursement

roduire des erreurs

Durabilité

4. La banque ne va pas oublier combien d'argent j'ai

Ni combien je leur dois…

La base de données de la banque

La cliente

Nom: Julie

Adresse: 1, rue Dugommier

Ville: Paris

Age: 22

Client	Nom	Adresse	Ville	Age
	Julie	1 rue Dugommier	Paris	22

La base de données de la banque

La cliente

Nom: Julie

Adresse: 1, rue Dugommier

Ville: Paris

Age: 22

Client	Nom	Adresse	Ville	Age
	Julie	1 rue Dugommier	Paris	22
	Marc	2 rue Archange	Orsay	25

La base de données de la banque

La cliente

Nom: Julie

Adresse: 1, rue Dugommier

Ville: Paris

Age: 22

Problème!...

Client				
	Julie	1 rue Dugommier	Paris	22
	Marc	2 rue Archange	Orsay	25
	Julie	1 rue Dugommier	Paris	22

La base de données de la banque

La cliente

Nom: Julie

Adresse: 1, rue Dugommier

Ville: Paris

Age: 22

Client	Nom	Adresse	Ville	Age
	Julie	1 rue Dugommier	Paris	22
	Marc	2 rue Archange	Orsay	25
	Julie	1 rue Dugommier	Paris	22

La base de données de la banque

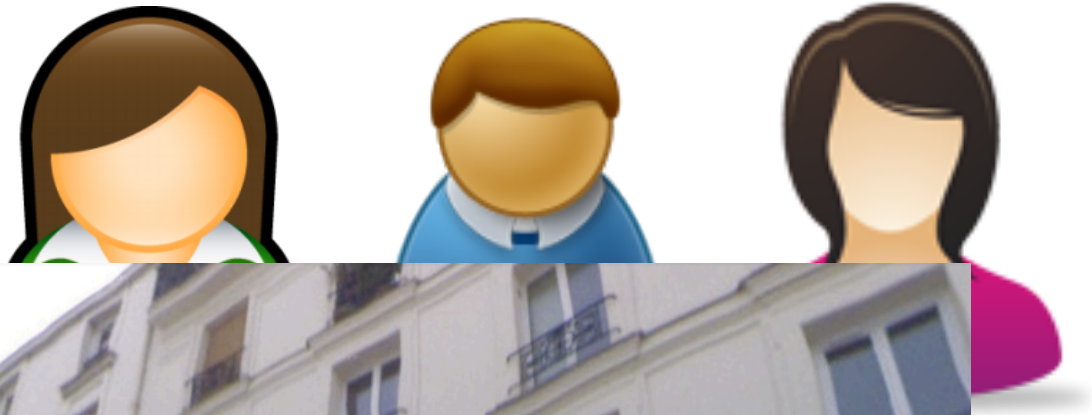
La cliente

Nom: Julie

Adresse: 1, rue Dugommier

Ville: P

Age: 22



Client

La base de données de la banque

La cliente

Nom: Julie

Adresse: 1, rue Dugommier

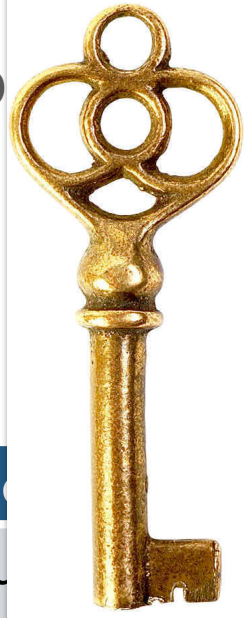
Ville: Paris

Age: 22

NumClient	Nom	Adresse	Ville	Age
1	Julie	1 rue Dugommier	Paris	22
2	Marc	2 rue Archange	Orsay	25
3	Julie	1 rue Dugommier	Paris	22

La base de données de la banque

La cliente
Nom: Julie
Adresse: 1, rue D
Ville: Paris
Age: 22

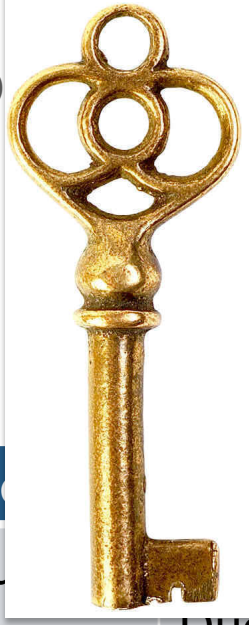


Clé primaire:
Connaître sa valeur permet d'identifier exactement un enregistrement

NumClient	N			Age
1	Ju	Dugommier		22
2	Marc	2 rue Archange	Orsay	25
3	Julie	1 rue Dugommier	Paris	22

La base de données de la banque

La cliente
Nom: Julie
Adresse: 1, rue D
Ville: Paris
Age: 22



Clé primaire:
Connaître sa valeur permet d'identifier exactement un enregistrement

NumClient	N		Age
1	Ju		22
2			
3			

C'est pour cela que les SAF appellent mon contrat LEO-EIT-GA-2011-HORS-5643

Les clients et les comptes

Clients

NumClient	Nom	Adresse	Ville	Age
1	Julie	1 rue Dugommier	Paris	22
2	Marc	2 rue Archange	Orsay	25
3	Julie	1 rue Dugommier	Paris	22

Comptes

NumCompte	Type	Découvert	NumClient
12345	Courant	1000	1

Les clients et les comptes

Clients

NumClient	Nom	Adresse	Ville	Age
1	Julie	1 rue Dugommier	Paris	22
2	Marc	2 rue Archange	Orsay	25
3	Julie	1 rue Dugommier	Paris	22

Comptes

NumCompte	Type	Découvert	NumClient
12345	Courant	1000	1
55555	Epargne-Logement	0	1

Les clients et les comptes

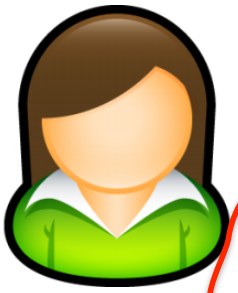
Clients

NumClient	Nom	Adresse	Ville	Age
1	Julie	1 rue Dugommier	Paris	22
2	Marc	2 rue Archange	Orsay	25
3	Julie	1 rue Dugommier	Paris	22

Comptes

NumCompte	Type	Découvert	NumClient
12345	Courant	1000	1
55555	Epargne-Logement	0	1
12000	Courant	2000	2

Les clients, les comptes et les transactions



Clients

NumClient	Nom	Adresse	Ville	Age
1	Julie	1 rue Dugommier	Paris	22
…	…	…	…	…

Comptes

NumCompte	Type	Découvert	NumClient
12345	Courant	1000	1
…	…	…	…

Transaction

NumCompte	Montant	Date	Info
12345	-40,00	5/10/11	Retrait
12345	+23,45	6/10/11	Remb. MAIF
12345	-300,00	7/10/11	Chaussures

Et s'il y a plusieurs bases distribuées?

=

Et s'il y a plusieurs bases distribuées?

Que peut-il arriver?

- Des ordinateurs peuvent tomber en panne
- Des messages peuvent se perdre → réseau partitionné

Théorème (Brewster / "CAP", 2000-2002)
Dans un système distribué, on ne peut pas avoir à la fois

1. Consistence: toutes les machines ont la même idée de l'état du système
2. Disponibilité: si on essaie de faire une opération, soit ça marche, soit on est informé que cela n'a pas marché
3. Résistance au partitionnement: cela continue à marcher même si des messages sont perdus

A conduit à des modèles de consistence "soft" (BASE: Basically Available Soft State with Eventual Consistency)
et des cauchemars aux fournisseurs d'infrastructure (PaaS)

2

Les reqûetes

Anne se demande: où habite Marc?

NumClient	Nom	Adresse	Ville	Age
1	Julie	1 rue Dugommier	Paris	22
2	Marc	2 rue Archange	Orsay	25
3	Julie	1 rue Dugommier	Paris	22

Marc est un Client de la banque. Cherchons dans la table des Clients.

Quand je trouve le nom Marc, j'aurai son adresse.

```
select Adresse
from Clients
where Nom="Marc"
```

Anne se demande: combien Marc a dans son compte?

Clients

NumClient	Nom	Adresse	Ville	Age
1	Julie	1 rue Dugommier	Paris	22
2	Marc	2 rue Archange	Orsay	25
3	Julie	1 rue Dugommier	Paris	22

Comptes

NumCompte	Type	Montant	NumClient
24126	Courant	2000	2

```
select Montant
from Clients, Comptes
where Nom="Marc" and
Clients.numClient=Comptes.NumClient
```

Anne se demande: quels clients ont des CEL?

Clients

NumClient	Nom	Adresse	Ville	Age
1	Julie	1 rue Dugommier	Paris	22
2	Marc	2 rue Archange	Orsay	25

Comptes

NumCompte	Type	Montant	NumClient
12345	Courant	1000	1
20000	CEL	2000	2

```
select Nom
from Clients, Comptes
where Type="CEL" and
Clients.numClient=Comptes.NumClient
```

Anne se demande: quels clients n'ont pas de CEL?

Clients

NumClient	Nom	Adresse	Ville	Age
1	Julie	1 rue Dugommier	Paris	22
2	Marc	2 rue Archange	Orsay	25

Comptes

NumCompte	Type	Montant	NumClient
12345	Courant	1000	1
20000	CEL	2000	2

```
select Nom
from Clients
where NumClient not in
(select NumClient
 from Comptes
 where Type="CEL")
```

Les requêtes et leurs problèmes

Inclusion de reqûetes: est-ce que pour toute base de données D,
Q(D) est inclus dans Q'(D)?

Q = "les clients Parisiens de moins de 30 ans"

Q' = "les clients Parisiens"

Décidable, NP-complet pour des requêtes conjonctives (Chandra et Merlin, 1977)

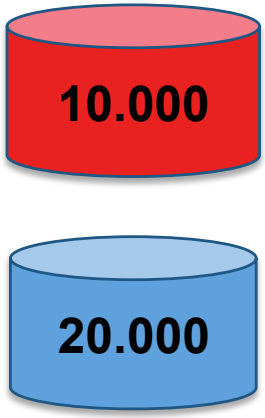
Indécidable pour des requêtes en algèbre relationelle (requêtes conjonctives + union + négation = FOL)

Essentiel pour: optimisation, contrôle d'accès, …

Equivalence de requêtes: est-ce que pour toute base de données D,
Q(D)=Q'(D) ?

Re-écriture de reqûetes à l'aide de vues:
si je connais les résultats de Q1, Q2, …, Qn,
est-ce que j'ai le moyen de calculer Q?

Comment calculer la fortune de Marc? (1)



Clients

NumClient	Nom	Adresse	Ville	Age
…	…	…	…	..

Comptes

NumCompte	Type	Montant	NumClient
24126	Courant	2000	2

200.000.000 opérations

Pour chaque client
Pour chaque compte
Si le Nom est "Marc"
Alors Si même NumClient
Alors retourner Montant

ct Montant
Clients, Comptes
where Nom="Marc" and
Clients.NumClient=
Comptes.NumClient

Comment calculer la fortune de Marc? (2)

Clients

NumClient	Nom	Adresse	Ville	Age
…	…	…	…	..

Comptes

NumCompte	Type	Montant	NumClient
24126	Courant	2000	2

Pour chaque client

Si le nom est "Marc"

Alors pour chaque com

Si même Nur

Alors retourner Montant

30.000 opérations maximum

ect Montant

m Clients, Comptes

ere Nom="Marc" and

ents.NumClient=

Comptes.NumClient

Combien ont coûté les chaussures de Julie?

Clients

NumClient	Nom	Adresse	Ville	Age
…	…	…	…	..

Comptes

NumCompte	Type	Montant	NumClient
…	…	…	…

Trois relations: Clients, Comptes, Transactions…
Optimisation de requête sur N relations:
$(N-1)*2^{N-2}$ (Ono and Lohman, 1990)

```
select Transaction.Montant
from Clients, Comptes, Transactions
where Client.Nom="Julie" and
    Clients.NumClient=Comptes.NumClient and
    Comptes.NumCompte=Transactions.NumCompte
```

Ça fait donc tout, une base de données?

Que fait-elle d'autre?

•Mises à jour: rajouter un utilisateur, mettre à jour son adresse, fermer un compte

•*Triggers* (évènement-condition-action):

- Lorsqu'on essaie de faire un paiement
- Et que le montant restant dans le compte irait en dessous du découvert autorisé
- Faire réfuser le paiement, envoyer une lettre au client…

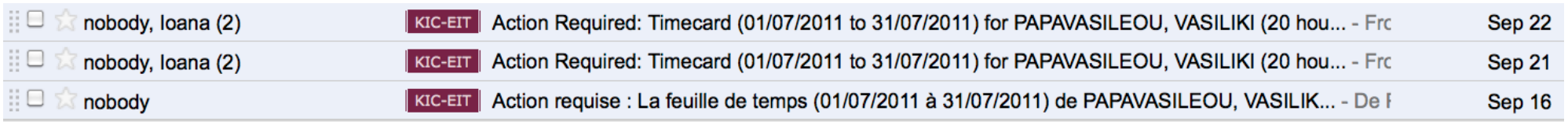
Evènement
Condition
Action

Elle ne fait pas tout.

- Les langages courants aujourd'hui permettent de faire des choses "simples"

(bien plus simples qu'un langage de programmation)

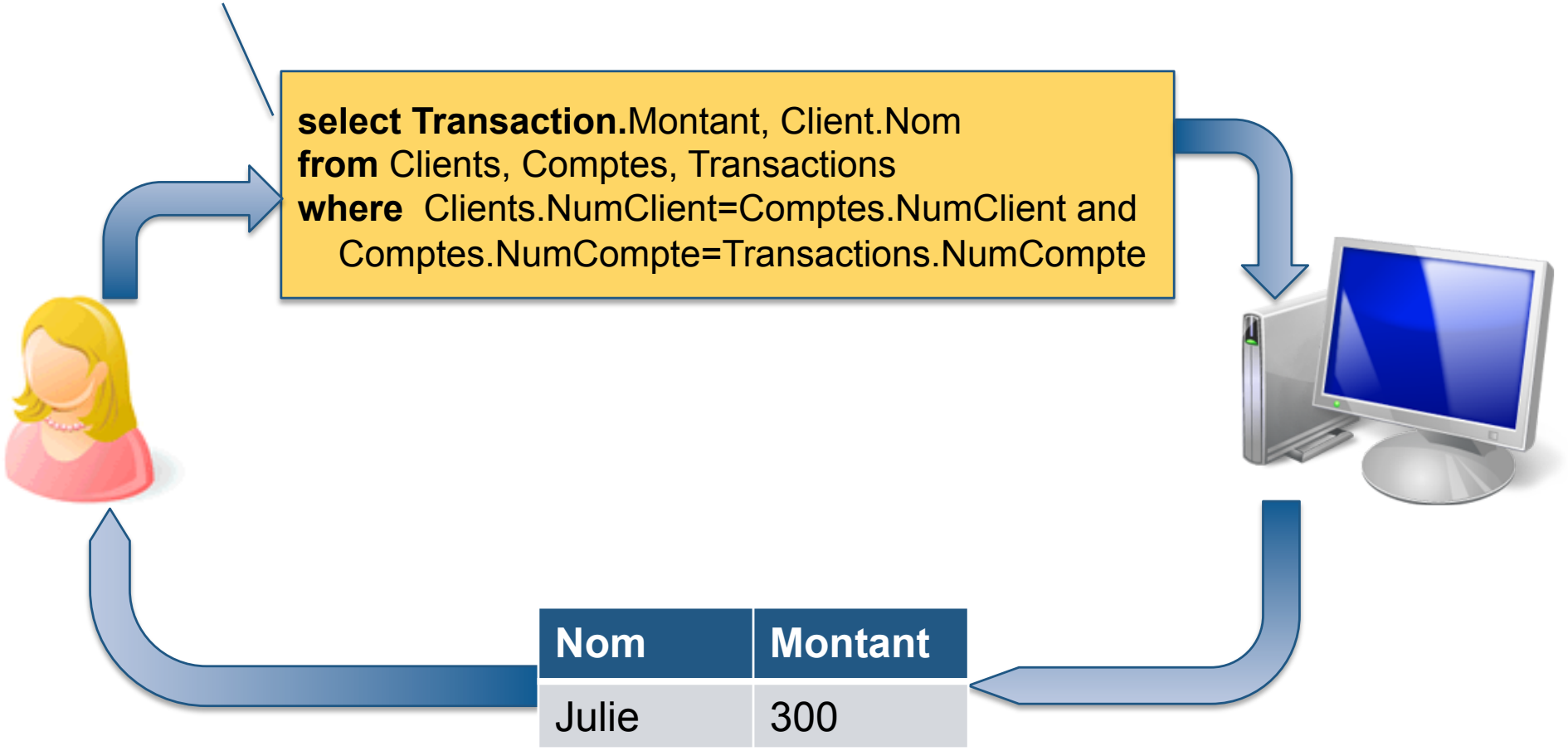
- Les applications sont faites souvent avec des bouts de programme autour de la base de données… Et ça donne ça:



3

Des bases de données au Web

Anne, son ordinateur, sa base de données

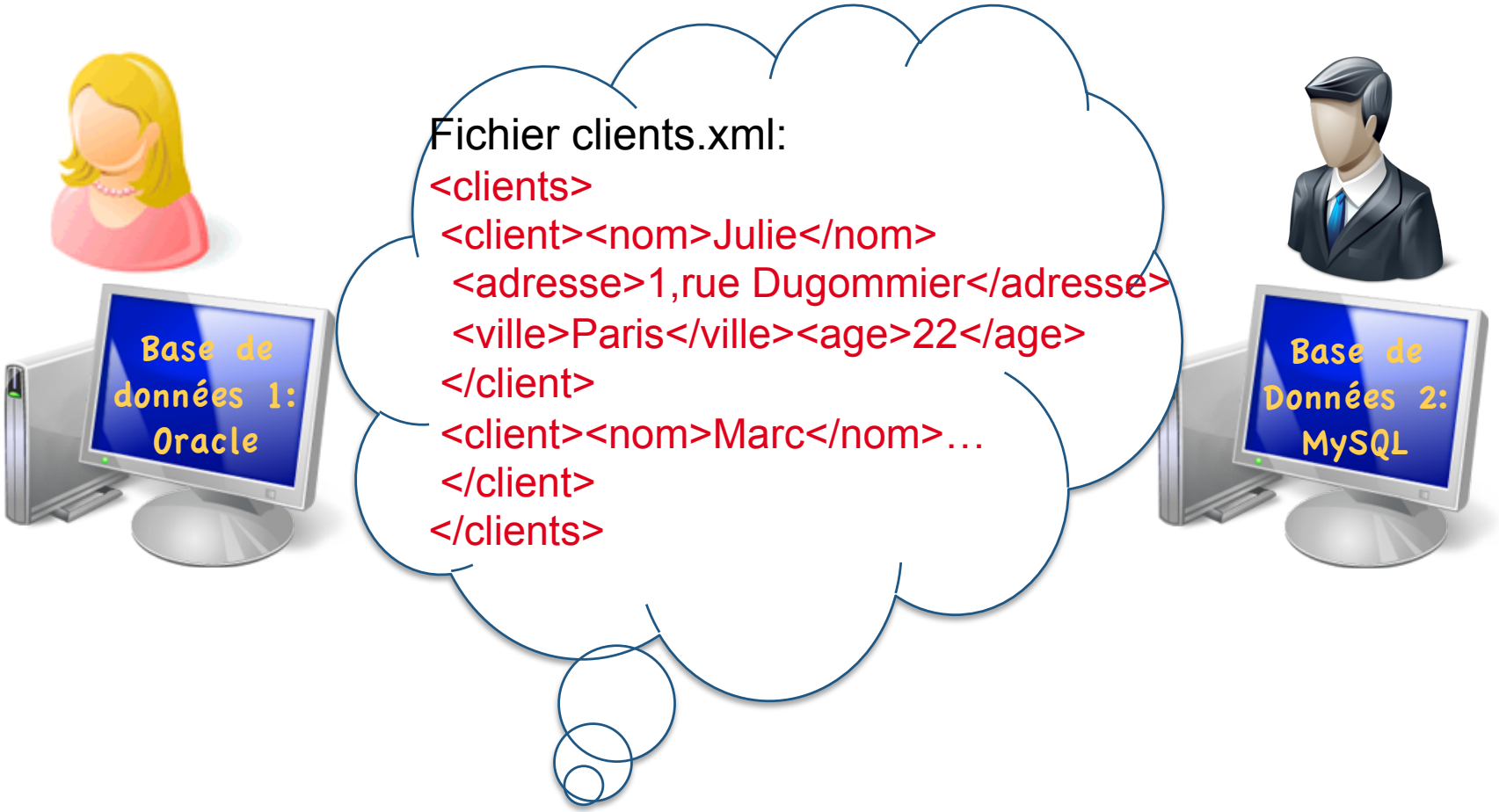


La banque d'Anne achète une autre banque

Comment s'échanger les données?

La banque d'Anne achète une autre banque

Comment s'échanger les données? Arrive XML (eXtensible Markup Language, WWW, 1998)



La banque d'Anne achète une autre banque

Comment s'échanger les données? Arrive XML (eXtensible Markup Language, WWW, 1998)

Source de : http://www.inria.fr/

```
'/institut/relations-internationales"><span>Relations internationales</span></a>

    <ul>
                <li><a href="/institut/relations-internationales/mot-d-helene-kirchner">Mot d'Hélène Kirchner</a></li>
                <li><a href="/institut/relations-internationales/partenariats-strategiques2">Partenariats stratégiques</a><
                <li><a href="/institut/relations-internationales/actions-dans-le-monde">Actions dans le monde</a></li>
                <li><a href="/institut/relations-internationales/appels-a-projets">Appels à projets</a></li>
                <li><a href="/institut/relations-internationales/contacts">Contacts</a></li>
        </ul>
                </li>
                        <li><a href="/institut/partenariats"><span>Partenariats</span></a>

    <ul>
                <li><a href=                                                    es</a></li>
                <li><a href=                                                    ls</a></li>
                <li><a href="/institut/partenariats/partenariats-europeens">Partenariats européens</a></li>
        </ul>
                </li>
                        <li><a href="/institut/recrutement-metiers"><span>Recrutement &amp;amp; métiers</span></a>

    <ul>
                <li><a href="/institut/recrutement-metiers/mot-de-muriel-sinanides">Mot de Muriel Sinanidès</a></li>
                <li><a href="/institut/recrutement-metiers/diversite-de-nos-metiers">Diversité de nos métiers</a></li>
                <li><a href="/institut/recrutement-metiers/nous-rejoindre">Nous rejoindre</a></li>
                <li><a href="/institut/recrutement-metiers/offres">Offres</a></li>
        </ul>
                </li>
            </ul>
```

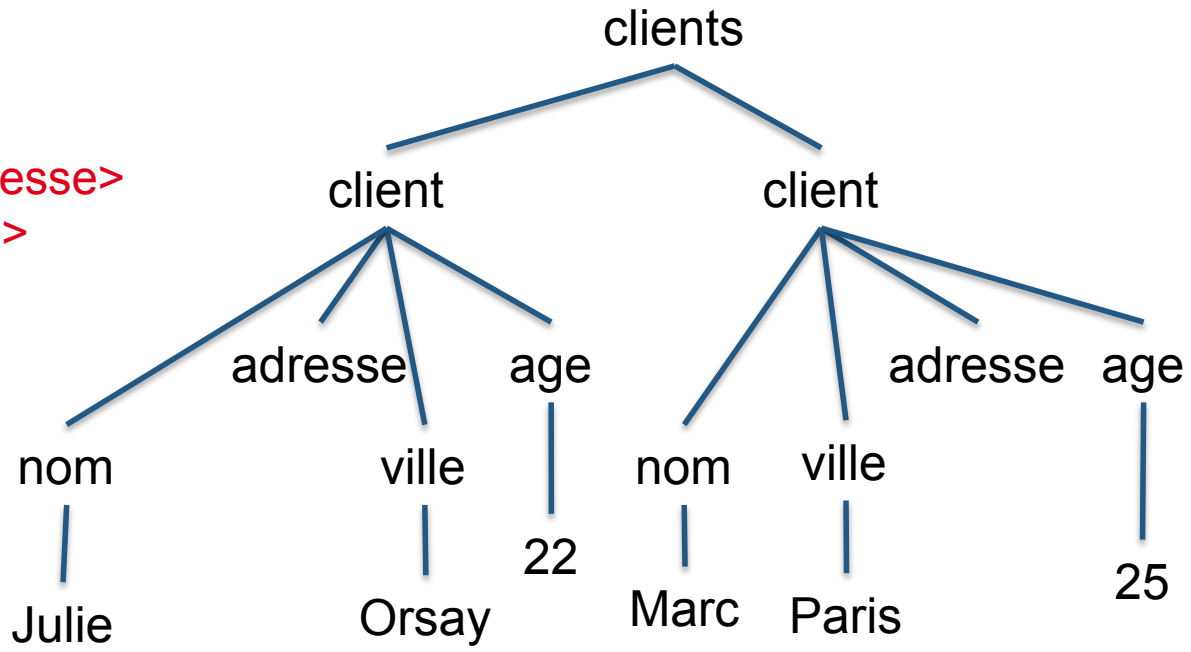
XML est devenu le langage du Web

XML: ce sont des arbres

Comment s'échanger les données? Arrive XML (eXtensible Markup Language, WWW, 1998)

Fichier clients.xml:

```
<clients>
 <client><nom>Julie</nom>
  <adresse>1,rue Dugommier</adresse>
  <ville>Paris</ville><age>22</age>
 </client>
 <client><nom>Marc</nom>…
 </client>
</clients>
```



Arbres XML = relations avec des contraintes d'intégrité

Relations:

- Noeud(idParent, idEnfant, ordreEnfant)
- Valeur(idNoeud, texte)

Contraintes d'intégrité:

- idNoeud est une clé pour Valeur
- idEnfant est une clé pour Noeud
- idEnfant détermine idParent
- idEnfant détermine ordreEnfant

Inclusion, équivalence de requêtes XML = relationnel + contraintes d'intégrité.

On peut rajouter des schémas XML…

Florescu (INRIA→Oracle), Levy (U. Washington→Google), Suciu (AT&T → U. Washington), Deutsch (UPenn → UCSD), Bidoit, Colazzo, Manolescu et bien d'autres

XML: cela ne s'arrête pas aux pages Web

Services Web: appels à une machine distante

- Requête = XML
- Réponse = XML
- Mise en oeuvre immédiate dès que vous avez un site Web qui tourne

Plein de problèmes intéressants d'échange, typage, coordination, synchronisation, vérification… Fortesse, Mexico, Dahu, WebDam, P. Dague

XML: et il y en a?

Il y en a *beaucoup*:

- Quasiment tous documents Web
- Tous les éditeurs type Office exportent en XML
- Fichiers de configuration d'autres logiciels
- Dans JPEG il y a du XML
- Essayez cela à la maison: recherche fichiers → par type → XML

Langage de requêtes XML: plus complexe (XQuery Turing complet)

On travaille sur des sous-langages (motifs d'arbre,
XQuery conjonctif, XQuery conjonctif + nœuds optionnels…)
Difficulté: identité des nœuds
Certains commencent à regarder JSON (XML sans identité)

Grande taille → optimisation

4

Le Web sémantique

Vision du Web
"compréhensible par les machines"

A quoi ça sert?

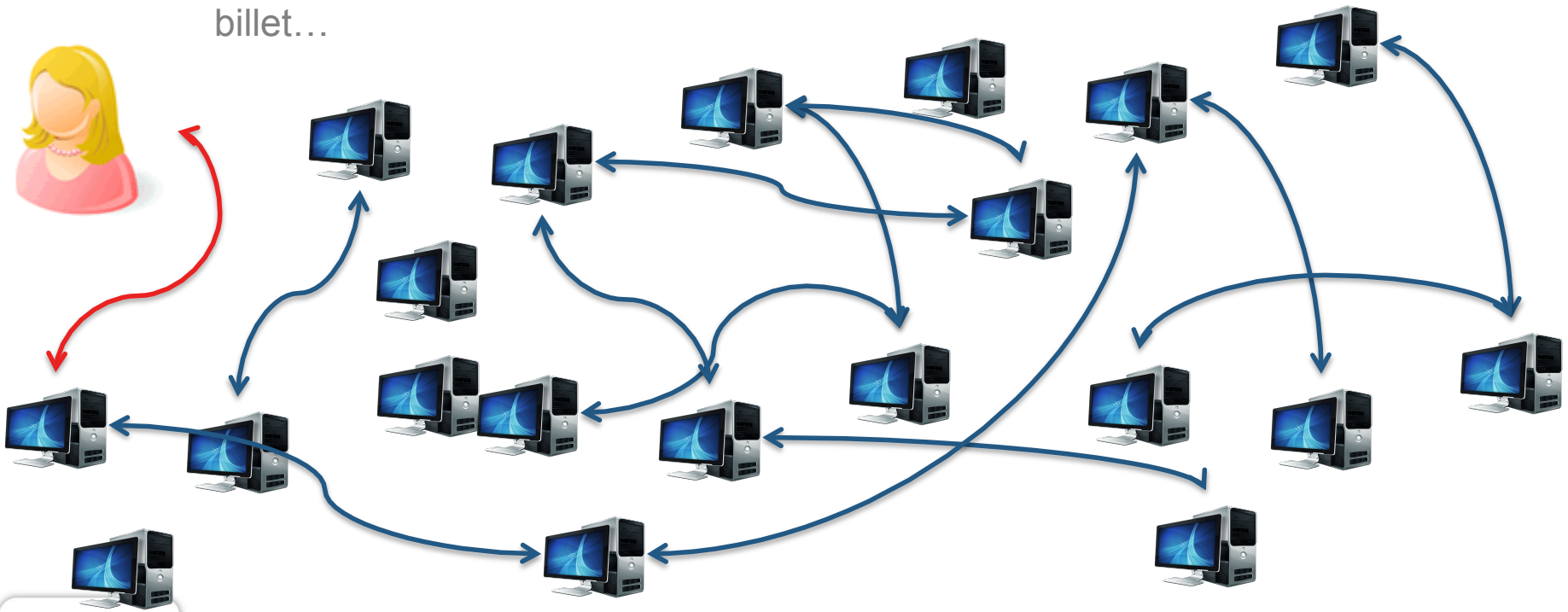
A rendre la recherche d'information intelligente et efficace

A combiner les sources d'information

hotel…

location skis…

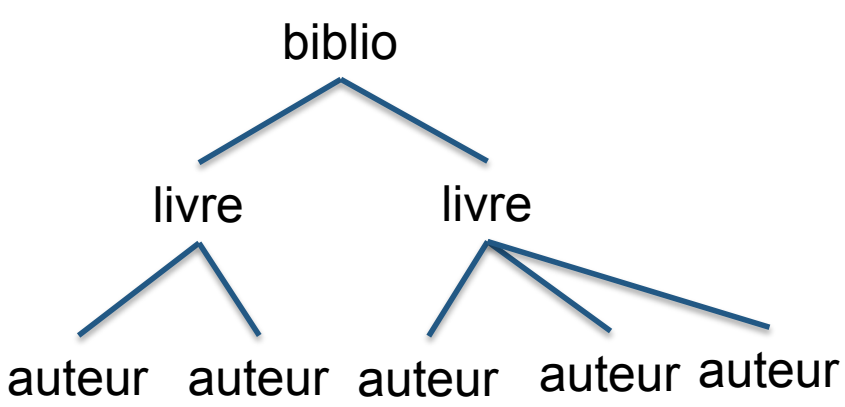
billet…

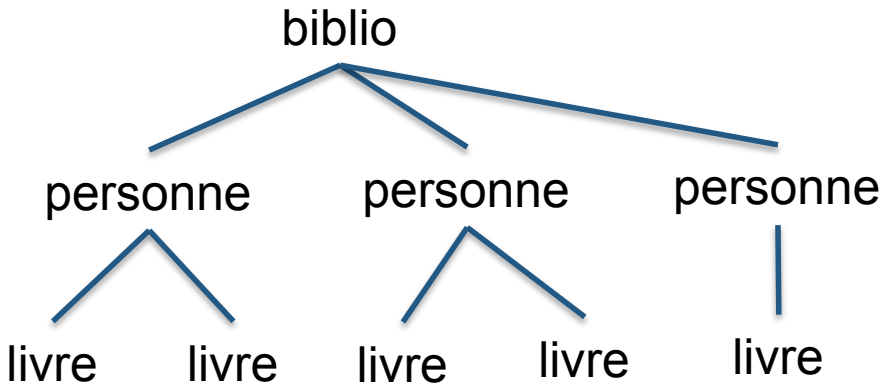


Vision du Web
"compréhensible par les machines"

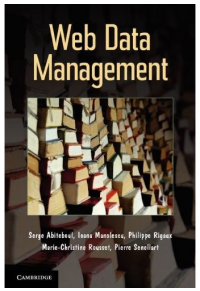
XML est organisé par documents

- Une racine, une seule façon de structurer
- Livres par auteur ou auteurs par livre?





Pas la même requête pour trouver



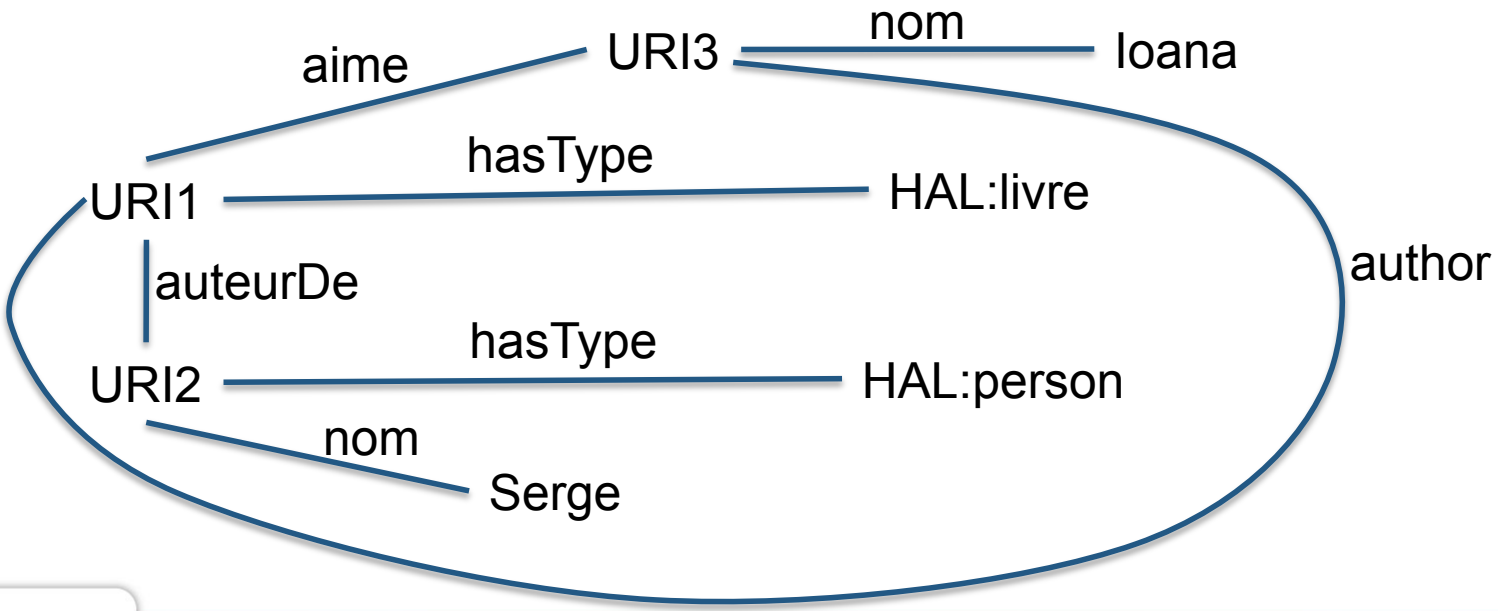
L'auteur du document *a raison.*

Web compréhensible par les machines

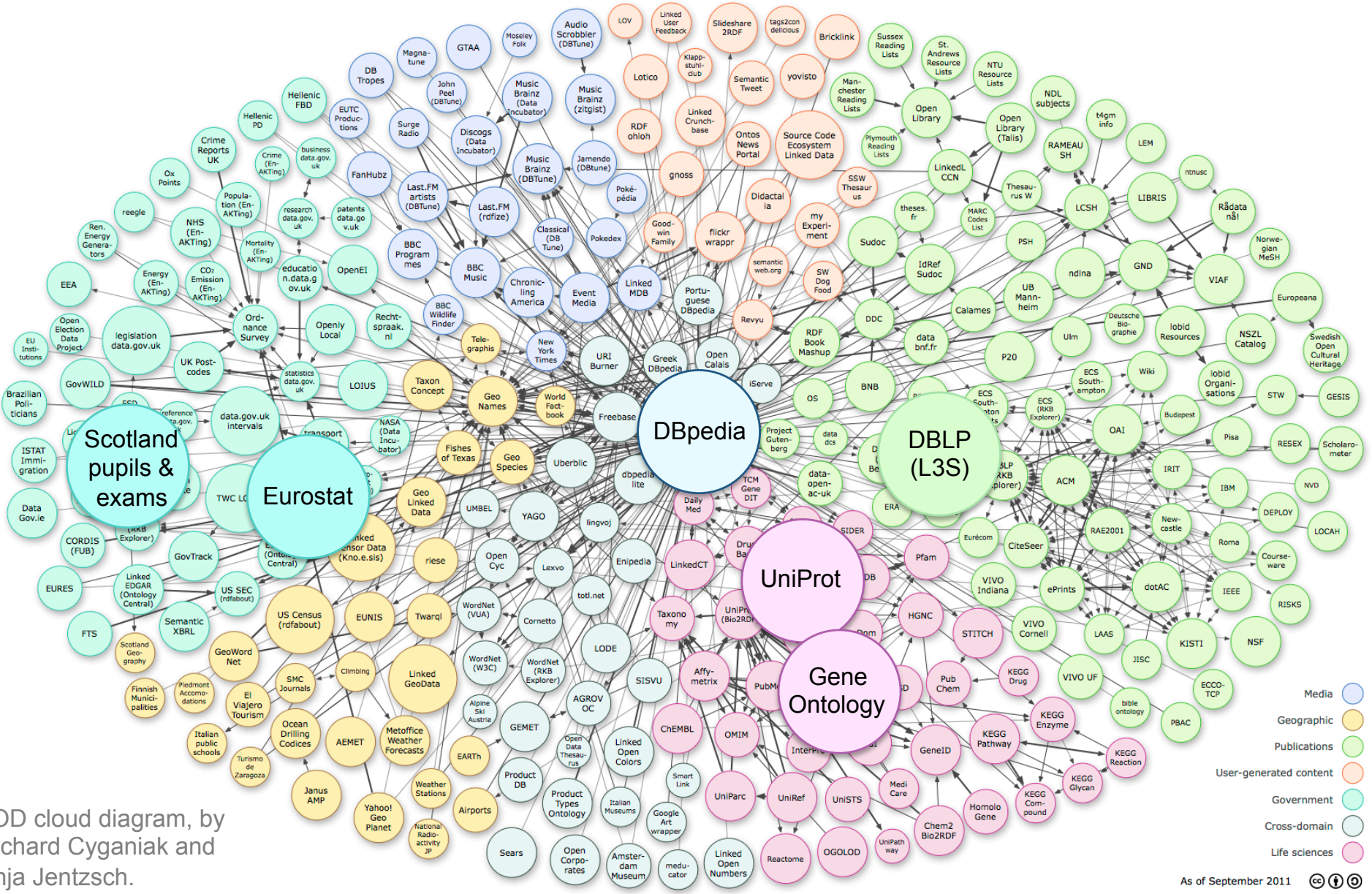
Arrive RDF: *Resource Description Framework*

Philosophie:

- Il y a des entités, qui ont des propriétés.
- Une entité a un identifiant unique (*Universal Resource Identifier*, URI)
- Une propriété a un nom.
- Une propriété d'une entité est soit une entité, soit une valeur



Grandes sources de données RDF: Linked Open Data

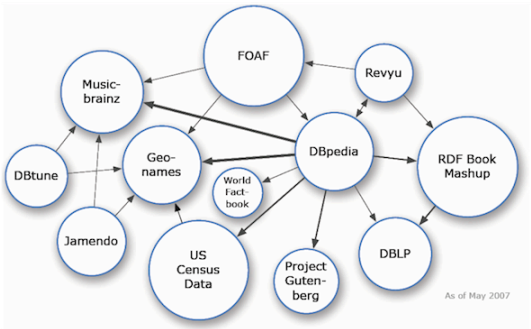


LOD cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

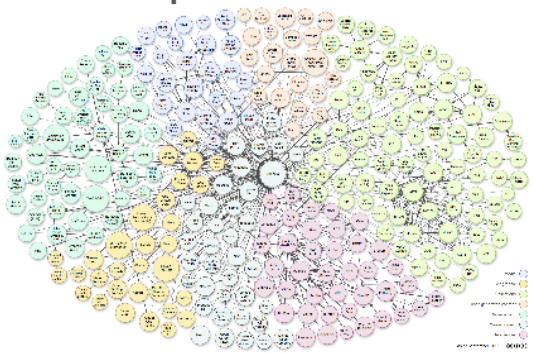
RDF: qu'y a-t-il à faire

1. Optimisation à très grande échelle
2. Intéropérabilité:
   - Comment trouver les clés?
   - Comment identifier les doublons?
   - Comment reconcilier les sources?
3. Comment fabriquer du Open Data?
4. Comment gérer la dynamique des sources?

Linked Open Data en 2005



Linked Open Data en 2011



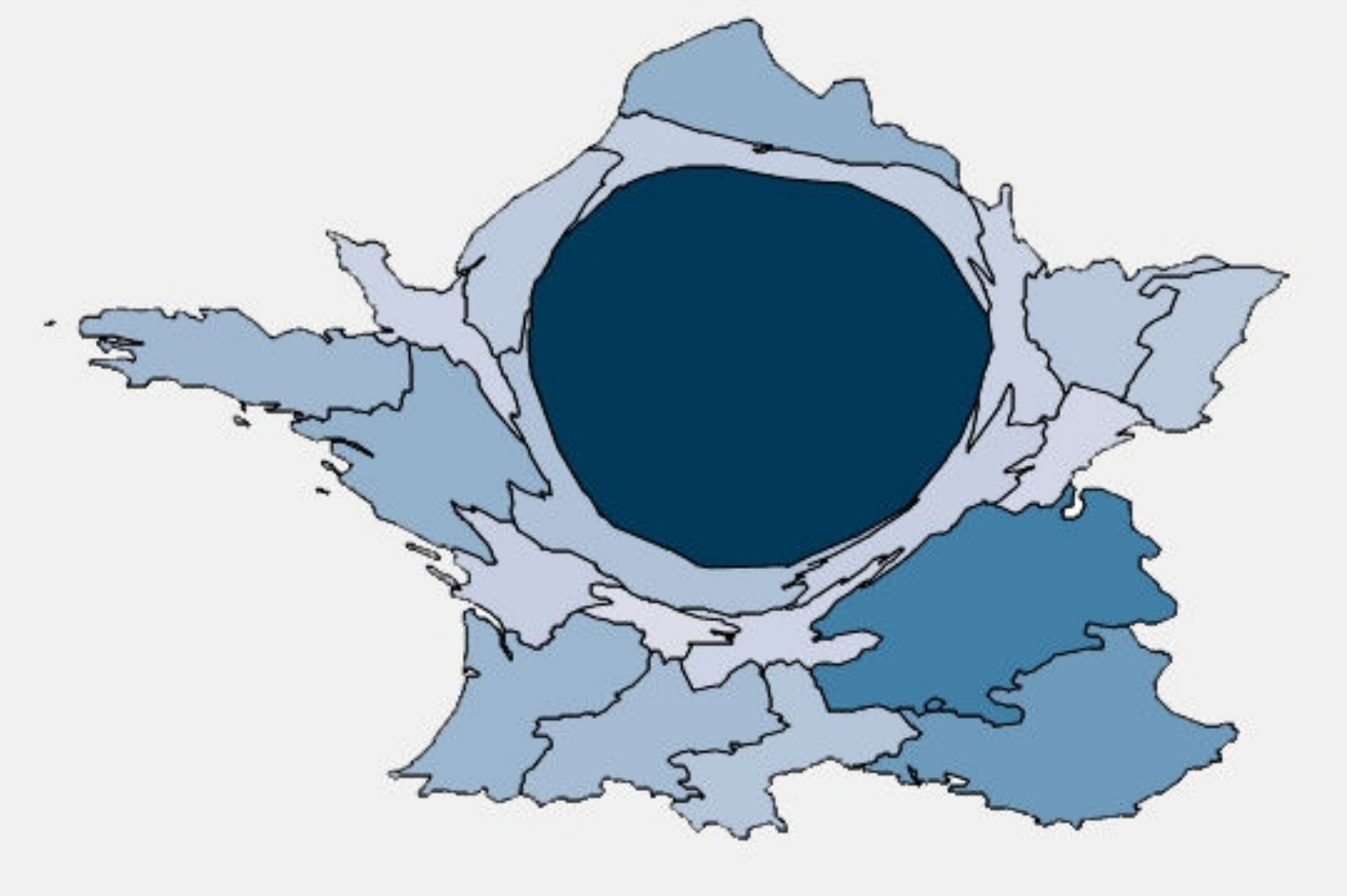
Très nombreuses équipes dans le monde dont LEO (C. Reynaud, F. Sais, N. Pernelle)

5

Un peu de Linked OpenData français

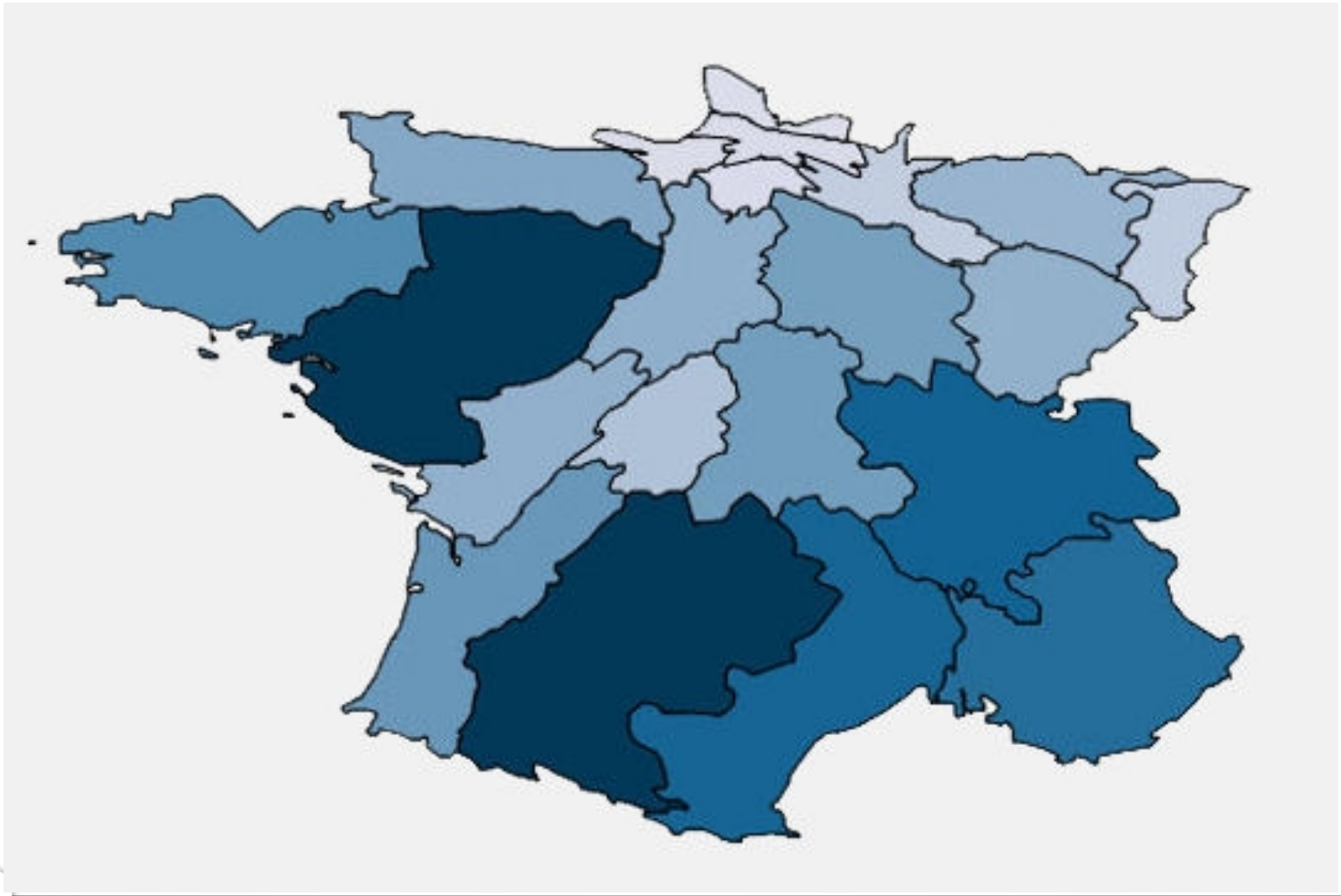
A partir d'Etalab (FR)

PIB par région (Le Journal Du Net, 07/09/2011)



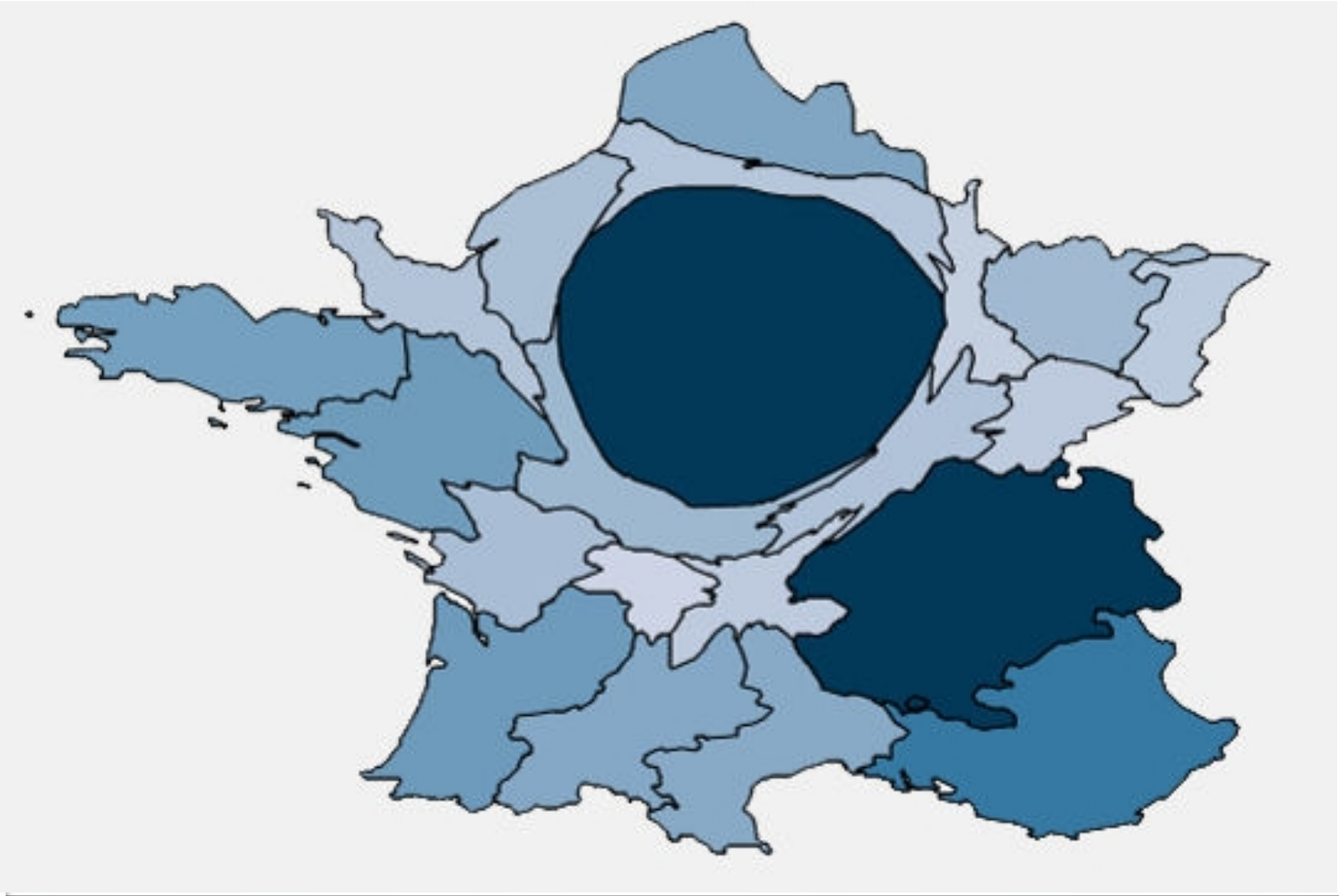
A partir d'Etalab (FR)

Agriculture bio par région (Le Journal Du Net, 07/09/2011)



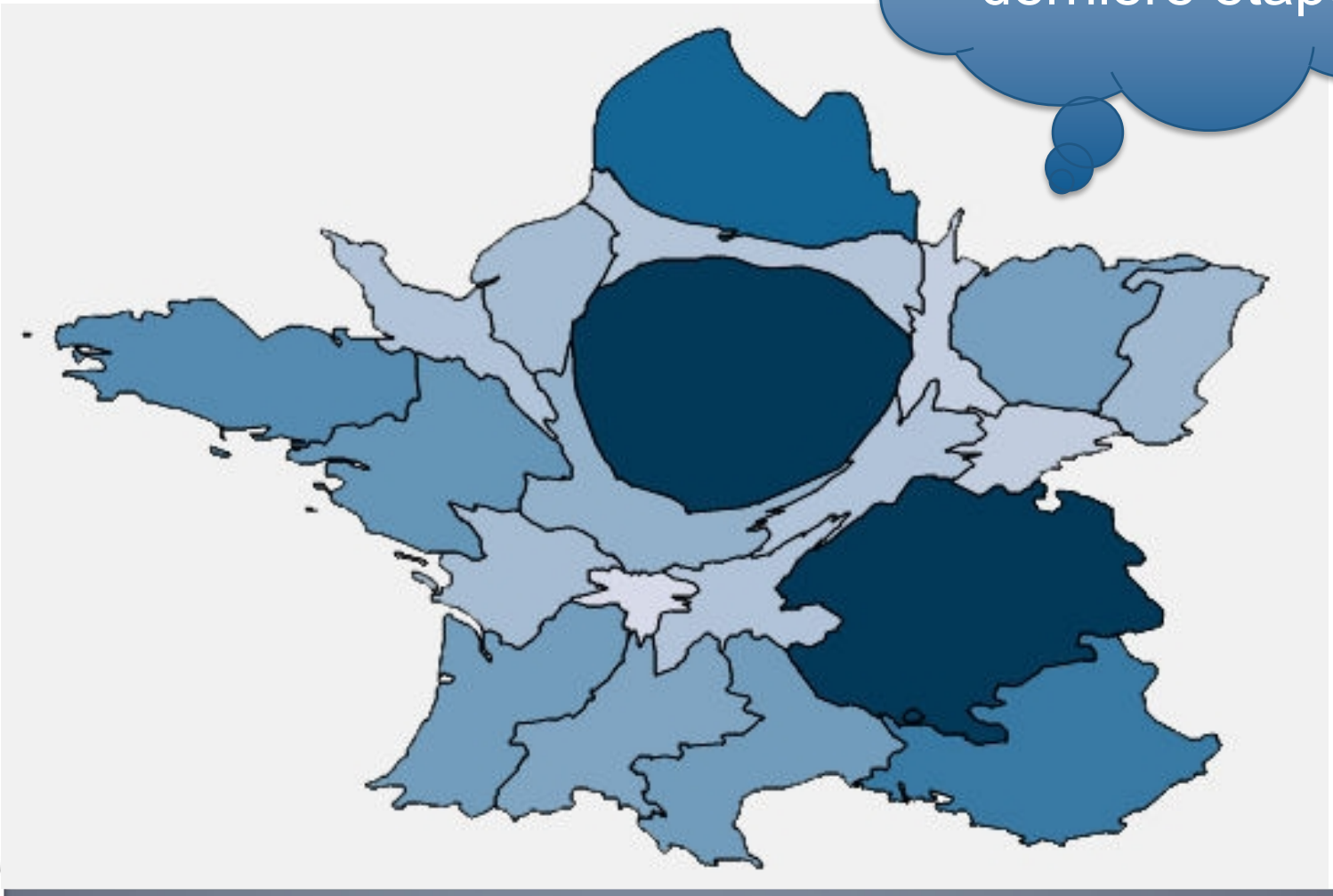
A partir d'Etalab (FR)

Cinémas par région (Le Journal Du Net, 07/09/2011)



A partir d'Etalab (FR)

Boulangeries par région (Le Journal Du Net, 07



Dessiner est seulement la dernière étape!

7

Conclusion

Data rocks

merci

LIEU
LOCALISATION

www.nomdedomaine.com