Hétérogénéité des données: enjeux et solutions

J. Villerd, F. Brun, O. Jolys, P. Métais, N. Munier-Jolain, O. Pointurier, F. Vuillemin

CoSAC

Contexte et enjeux

Déluge de données

- des données massives
- dans des formats différents (Cosac)

problème de l'intégration et de l'interopérabilité de ces données

La diversité dans Cosac

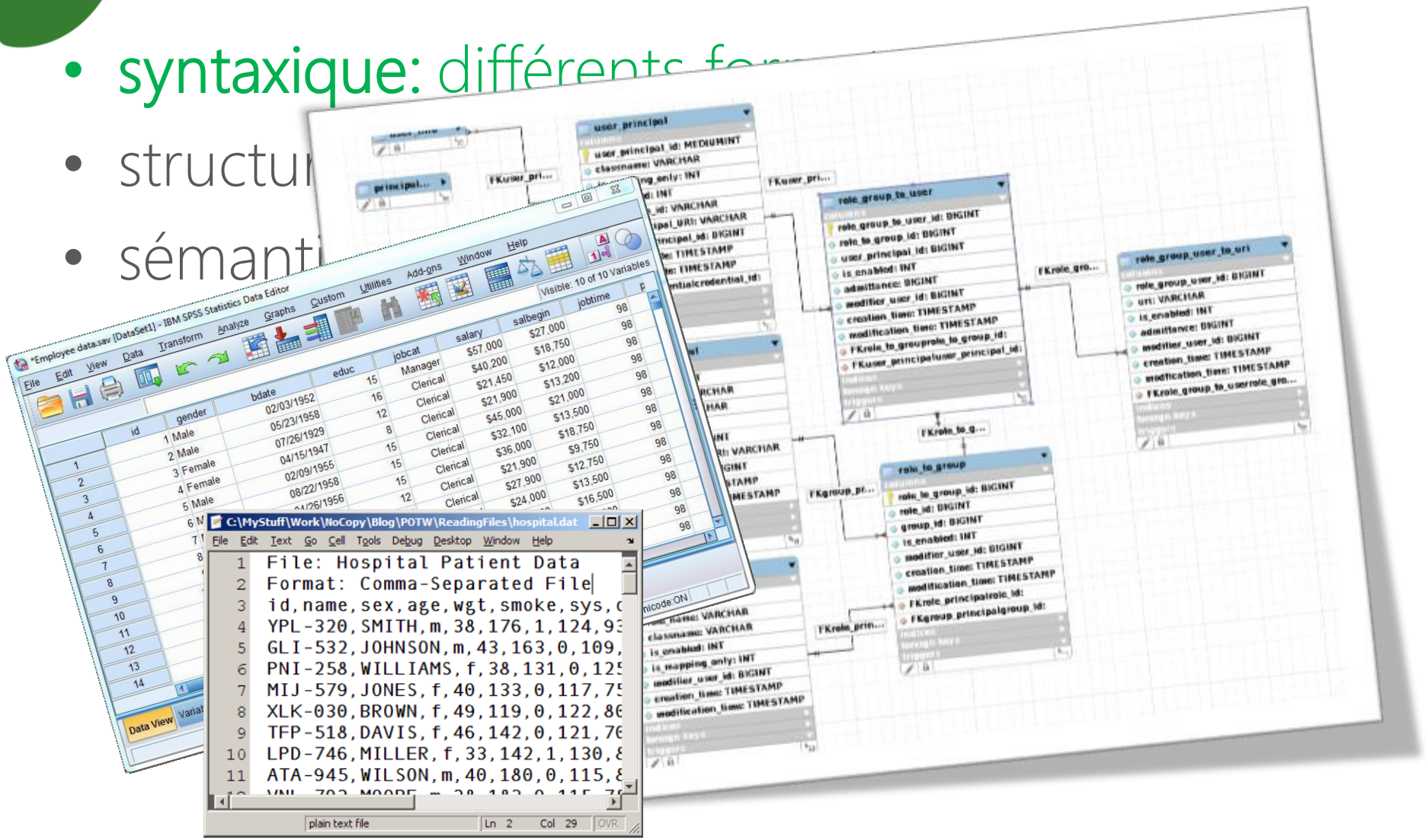
a pour conséquence une hétérogénéité dans les données manipulées

Niveaux d'hétérogénéité

- syntaxique
- structurel
- sémantique

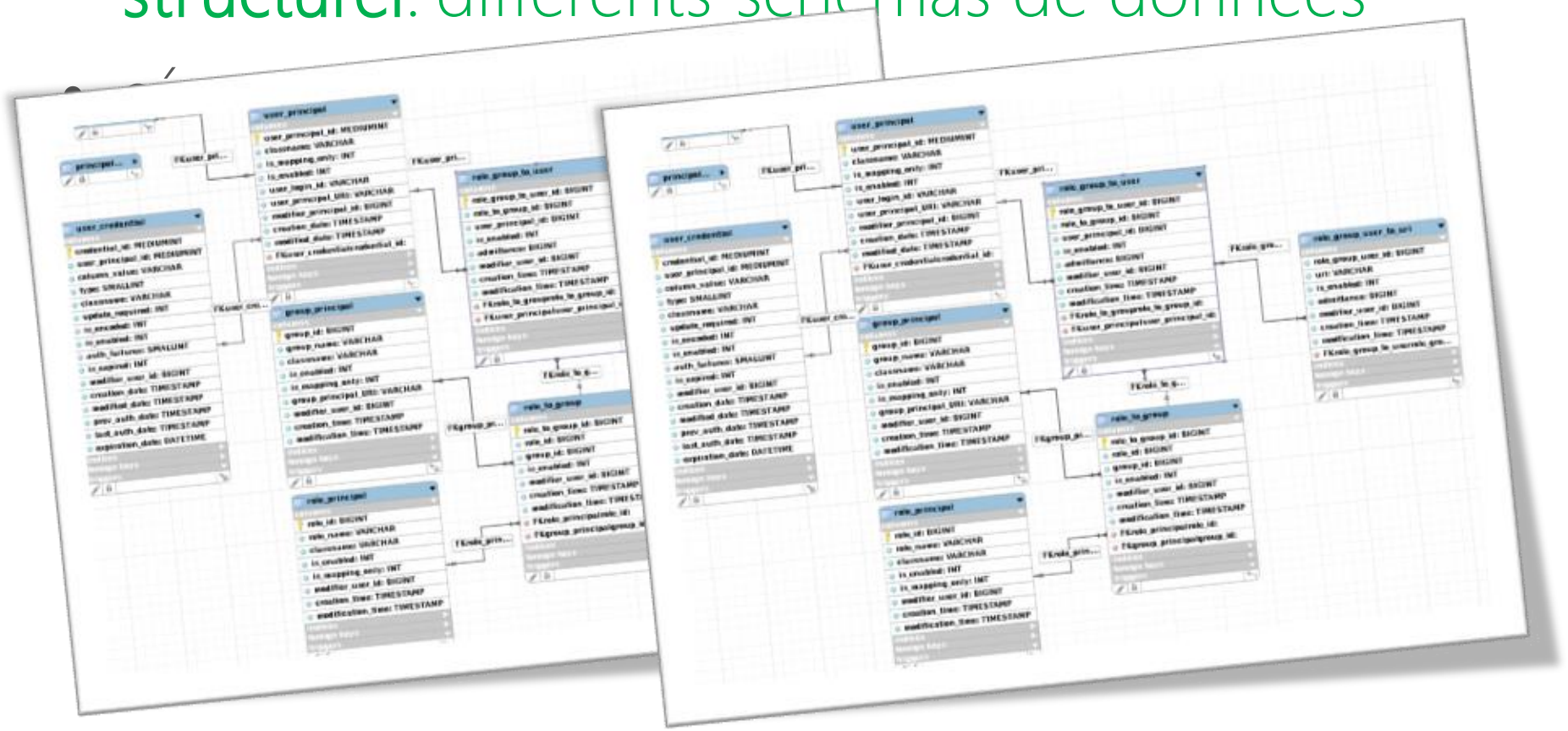
Niveaux d'hétérogénéité

- **syntaxique:** différents fo
- structur
- sémanti



Niveaux d'hétérogénéité

- syntaxique
- **structurel**: différents schémas de données

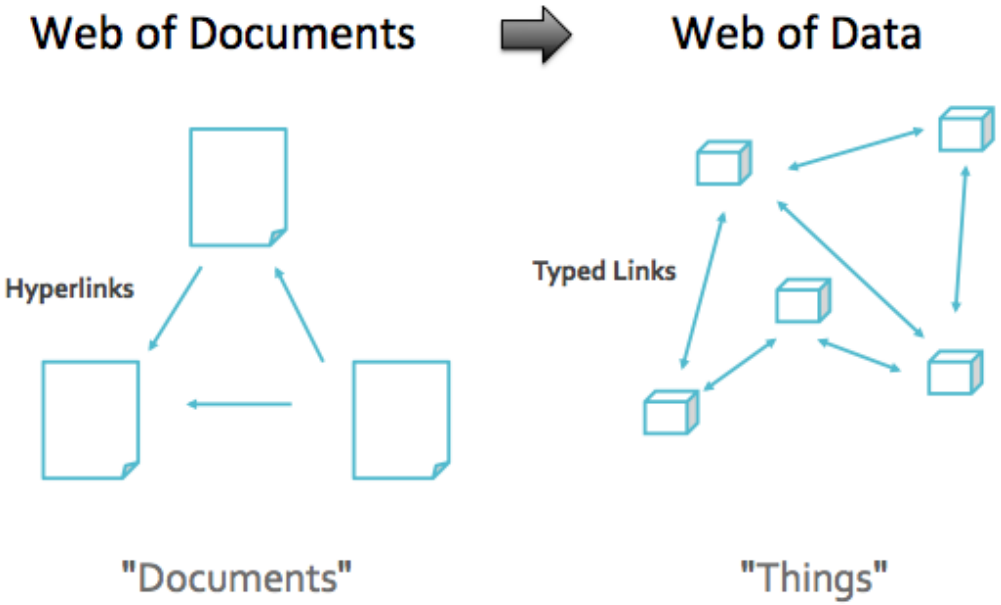


Niveaux d'hétérogénéité

- syntaxique
- structurel
- **sémantique**: absence de vocabulaire partagé
  - problème du nommage des concepts
  - plusieurs noms pour une même espèce
  - problème des traductions

L'idéal: les données liées

- données liées (*linked data*) = contenus annotés par des vocabulaires partagés (taxonomies, ontologies, etc.)
- en ligne, forment le web des données



Gestion des données dans le projet Cosac

CoSAC

Utilisation des données

- méta-analyse
- entrée de modèles
- analyse pour l'extraction de connaissances
- évaluation des outils et modèles

- constitution d'un groupe de travail
- organisation d'un journée

Solutions d'intégration

- solution lourde
  - un SI complet (BDD + référentiels) type Systerre ou Agrosyst
  - potentiellement coûteuse en temps
  - pérennité de l'intégration des données
- solution légère
  - un format structuré ad hoc type tableur
  - plus simple à mettre en œuvre
  - interopérabilité *a minima*

Choix pragmatique

- solution lourde trop longue pour être retenue dans le cadre interne du projet
- mais à envisager pour le stockage des données issues du projet ?
- la solution retenue est un fichier structuré (csv) utilisant un format d'identifiant unique et des référentiels permettant d'assurer l'interopérabilité

Interopérabilité Agrosyst-Florsys

- automatiser l'import de systèmes de culture depuis Agrosyst vers Florsys
- en cours de développement pour test sur les SdC du réseau Dephy
- permettra de simuler Florsys sur tout SdC décrit dans Agrosyst

Hétérogénéité des données

- problème non trivial même avec un nombre de fournisseurs de données restreint
- la mise en place d'une solution pérenne à long terme est longue et coûteuse
- un groupe de travail incluant les fournisseurs de données est indispensable
- pertinence d'un *data management plan* ?

solution légère retenue

couplage Agrosyst Florsys en cours

Réflexion en amont