Alimentation en ligne d'une base de données à partir d'un fichier Excel

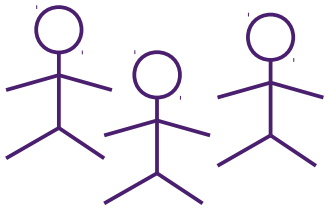
Marie-Claude QUIDOZ
CEFE - Montpellier

Juliette FABRE
OSU OREME - Montpellier

3èmes Rencontres R - Montpellier - 26/06/2014

Contexte

Producteurs données
& utilisateurs de R



Gestion des données
(saisie, traitement, visualisation ..)

Aucune contrainte
→ pas de typage
→ pas de structure
→ pas de contrôle

↓

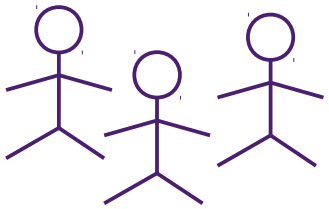
- Incohérences
- Fautes de frappe

PostgreSQL

Pérenniser les données
Offrir des outils de diffusion

Projet

Producteurs données
& utilisateurs de R



Gestion des données
(saisie, traitement, visualisation ..)

Aucune contrainte
→ pas de typage
→ pas de structure
→ pas de contrôle

- Incohérences
- Fautes de frappe

Insertion automatisée
Vérification des données

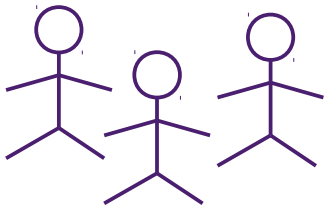
PostgreSQL

Pérenniser les données
Offrir des outils de diffusion

Ajout de deux contraintes

Producteurs données
& utilisateurs de R

Gestion des données
(saisie, traitement, visualisation ..)



Aucune contrainte
→ pas de typage
→ pas de structure
→ pas de contrôle

Pas de manipulation

1

- Incohérences
- Fautes de frappe

Insertion automatisée en ligne
Vérification des données

2

Pérenniser les données
Offrir des outils de diffusion

Fonctions génériques R

Script R de fonctions utilitaires et génériques

- Interrogation de bases de données (RPostgreSQL)
  - → Extraction de valeurs : comparaison données / BDD
  - → Ordination de jeux de données selon une table : formatage
  - → Alternative : RODBC

Fonctions génériques R

Script R de fonctions utilitaires et génériques

- Interrogation de bases de données (RPostgreSQL)
  - → Extraction de valeurs : comparaison données / BDD
  - → Ordination de jeux de données selon une table : formatage
  - → Alternative : RODBC

- Vérification de données
  - → Unicité de valeurs : clés primaires
  - → Valeurs manquantes : NOT NULL, clés étrangères
  - → Valeurs autorisées : liste, intervalle
  - → Format : numérique, date, datetime, nb de caractères, ...

Fonctions génériques R

Exemples :

- Les noms de stations existent-ils dans la base ?

```
db_station <- get_db_values(fields = 'station_name',
                            table = 'station',
                            schema  = 'my_schema')

check_belong(col = 'station', data = data, value_set = db_station)
```

- Les sample_id sont-ils fournis, uniques et de 30 caractères max ?

```
check_missing_values(col = 'sample_id', data = data)

check_unicity(col = 'sample_id', data = data)

check_nb_character(col = 'sample_id', data = data, nb_char = 30)
```

Format et règles de validation

- Définition d'un format standard pour les fichiers Excel

  → Pour chaque type de données

  → Définition des noms de colonnes, de feuilles

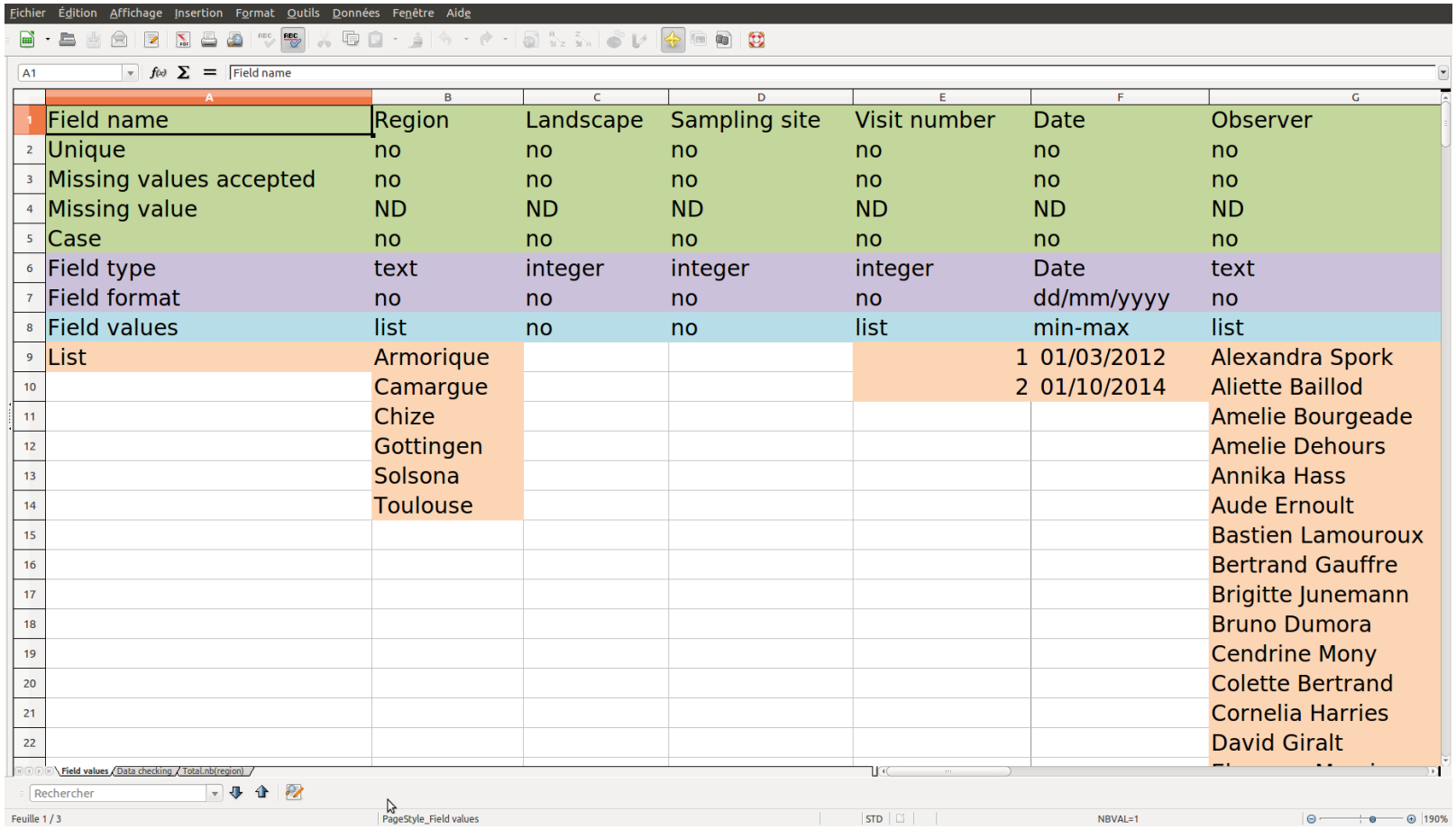
  → flexibilité possible : lignes / colonnes supplémentaires non prises en compte

Format et règles de validation

- Définition de règles de validation

  → Pour chaque colonne : type, format, valeurs, unicité, ...

  → Dans le fichier comme feuille de métadonnées



	A	B	C	D	E	F	G
1	Field name	Region	Landscape	Sampling site	Visit number	Date	Observer
2	Unique	no	no	no	no	no	no
3	Missing values accepted	no	no	no	no	no	no
4	Missing value	ND	ND	ND	ND	ND	ND
5	Case	no	no	no	no	no	no
6	Field type	text	integer	integer	integer	Date	text
7	Field format	no	no	no	no	dd/mm/yyyy	no
8	Field values	list	no	no	list	min-max	list
9	List	Armorique			1	01/03/2012	Alexandra Spork
10		Camargue			2	01/10/2014	Aliette Baillod
11		Chize					Amelie Bourgeade
12		Gottingen					Amelie Dehours
13		Solsona					Annika Hass
14		Toulouse					Aude Ernoult
15							Bastien Lamouroux
16							Bertrand Gauffre
17							Brigitte Junemann
18							Bruno Dumora
19							Cendrine Mony
20							Colette Bertrand
21							Cornelia Harries
22							David Giralt

Scripts R de traitement

Scripts R de traitement (pour chaque type de données)

- Import des fichiers avec XLConnect

  → multi-plateformes, gestion des caractères spéciaux, des noms de feuilles

  → loadWorkbook(), getSheets, readWorksheet()

- Vérification des données (fonctions génériques)

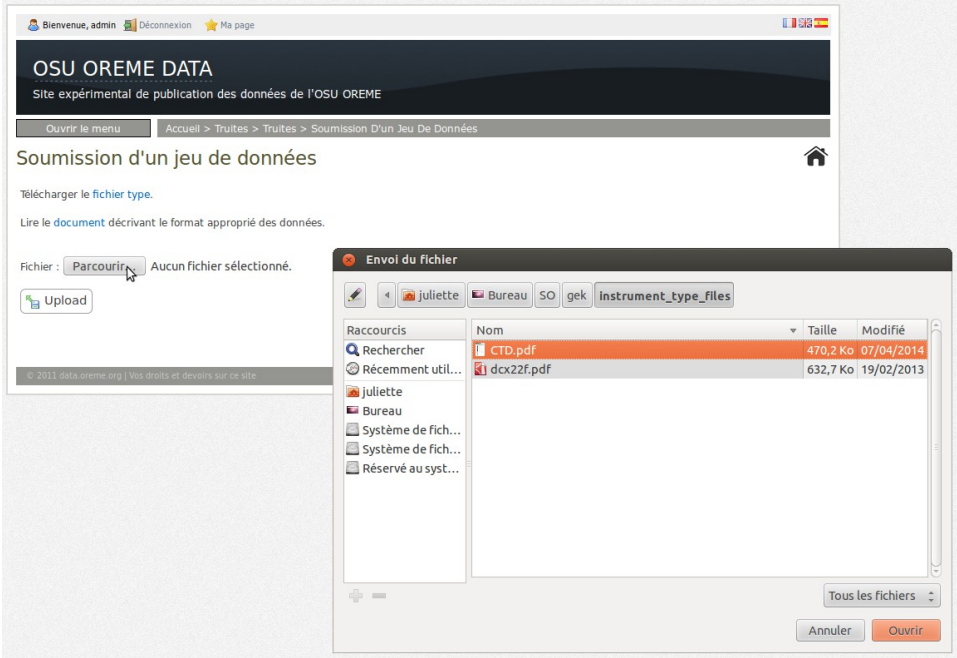
Formatage (ajout d'identifiants uniques, traitements, séparation en plusieurs tables, ...) et insertion

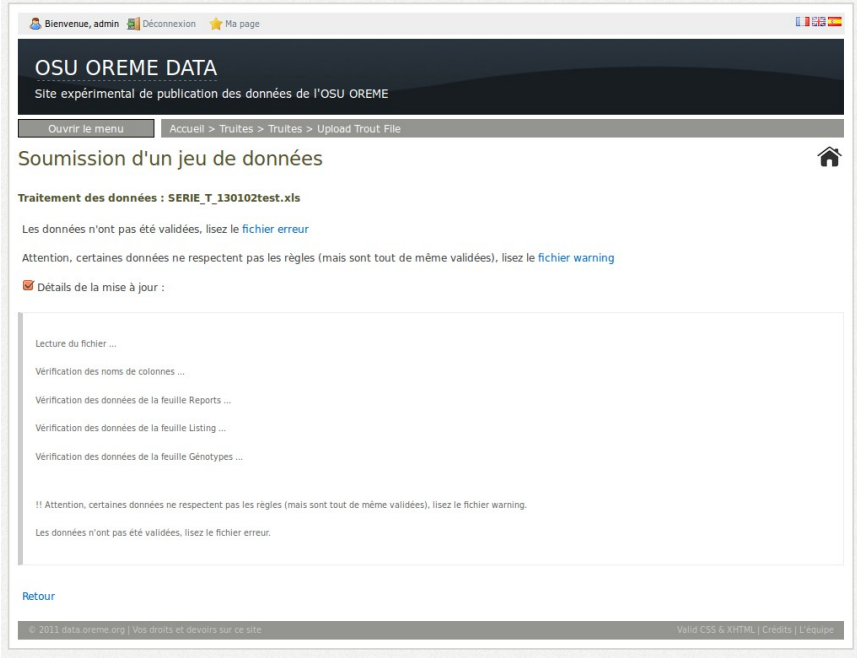
Ecriture des erreurs dans un fichier

Interface web

Interface web de soumission de fichier (PHP)

- Upload
- Exécution du script R
- Récupération et téléchargement des fichiers d'erreurs





Conclusions (1)

- Import des fichier Excel : XLConnect

  → Pb avec les fichiers volumineux (ex : 13Mo, 150.000 lignes)

  options(java.parameters = "-Xmx4096m") ne suffit pas toujours

  → Bug dans la lecture de champs heure

  → Alternative : xlsx

- Fichiers Libre Office (.ods)

  → librairies gnumeric, ROpenOffice, speedR

  → gnumeric, speedR : pas réussi à utiliser

  → ROpenOffice : moins performante ? Difficultés d'import

Conclusions (2)

- Réalisation dans différents environnements

• Debian GNU/Linux 7 (wheezy) • R 3.1 • PostgreSQL 9.2.8 • Apache 2.2.22 • PHP 5.3.1011	• Windows Server 2008R2 • R 3.1 • PostgreSQL 9.3.4 • Apache 2.4.9 • PHP 5.5.10

- Application web → gestion des droits nécessaire

  → Upload du fichier, lecture du fichier Excel, exécution du script R

- Généricité du script de vérification

  → difficultés à décrire formellement l'intégralité des règles de validation