Utilisation de technologies Open Source dans l'administration, la qualification et l'exploitation des données du Système d'Information sur l'Eau

Un zoom sur l'utilisation de PostgreSQL, PostGIS et PLR

Intervenants : Alexandre Liccardi (ONEMA)
Jean-Philippe Goyen (ONEMA)

Le Système d'information sur l'Eau

Un outil national coordonné par l'ONEMA

Le Système d'Information sur l'Eau (SIE) est un dispositif partenarial regroupant les principaux acteurs publics du domaine de l'eau , qui organise la collecte, le stockage, la valorisation et la diffusion des données sur l'eau, les milieux aquatiques et leurs usages.

- **2003** *Réseau national des données sur l'eau* issu de la loi sur l'eau de 1992
- **2006** Introduit dans le code de l'environnement par la loi sur l'eau et les milieux aquatiques
- **2009** Schéma National des Données sur l'Eau (SNDE)

Des dizaines de millions de données dans plus de 15 banques nationales de référence :

- ✓ séries longues de mesures
- ✓ produits de calculs, d'expertise ou d'évaluation
- ✓ référentiels, données géographiques
- ✓ données de rapportage européen

► Plus de 15 000 structures contributrices !

L'accès aux informations est garanti par la toile *EauFrance* (plus de 30 sites Web)

Le projet *Données pour la décision*

OpenSource & mise à disposition des données

Entre autres missions

- mise en qualité de l'information (en garantissant sa cohérence)
- accès à l'information par les différents publics
- analyse de ces informations
- aide à la décision technique, administrative ou économique (actions de restauration, de définition de programmes de mesures et du contrôle des usages de l'eau)

Mobilisation de technologies Open Source *pour :*

→ **extraire** des séries de données de fréquence de rafraîchissement et de volumétrie très variables
→ appuyer la **construction d'indicateurs** d'état des eaux et/ou de performance des politiques publiques
→ participer aux **stratégies de mobilisation et de réutilisation** des données en ligne (avec le pôle INSIDE)
→ (de manière plus générale) réaliser des **travaux de contrôle** d'intégrité des référentiels, de complétude des séries, identifier des **informations aberrantes ou illogiques**, vérifier la **cohérence** à des règles de gestion, **mettre en qualité** les données selon les précédents concepts

Données environnementales : besoins *inédits*

Des cas techniques et scientifiques bien particuliers…

Le nombre de pêcheurs ?

Le nombre moyen de civelles par kg de capture en 2007 ?

Le nombre de cours d'eau en Europe ?

(Comme autant d'employés de PME.)
Oui, facile.

(Toutes les données sont disponibles.)
Oui, facile.

(Talend doit bien savoir faire ça.)
Oui, facile.

Données environnementales : besoins *inédits*

Des cas techniques et scientifiques bien particuliers…

Le nombre de pêcheurs ?

Le nombre moyen de civelles par kg de capture en 2007 ?

Le nombre de cours d'eau en Europe ?

(Comme autant d'employés de PME.)
Oui, facile.

(Toutes les données sont disponibles.)
Oui, facile.

(Talend doit bien savoir faire ça.)
Oui, facile.

Caractéristiques non-additives

Complexité des changements d'échelle et des parcours de réseau

Historique technique et administratif

Sens métier à associer aux scénarios

(…)

Données environnementales : besoins *inédits*

Des cas techniques et scientifiques bien particuliers…

Le nombre de pêcheurs ?

Le nombre moyen de civelles par kg de capture en 2007 ?

Le nombre de cours d'eau en Europe ?

(Comme autant d'employés de PME.)
Oui, facile.

(Toutes les données sont disponibles.)
Oui, facile.

(Talend doit bien savoir faire ça.)
Oui, facile.

Caractéristiques non-additives

Historique technique et administratif

Complexité des changements d'échelle et des parcours de réseau

Sens métier à associer aux scénarios

(…)

Quelques exemples de développements informatiques nécessaires…

Parcours des réseaux hydrographiques ou scénarios de décision

Priorisation de l'action territoriale

Mise en qualité des données

Description de la relation EQ/MdO/Station dans Rapportages DCE

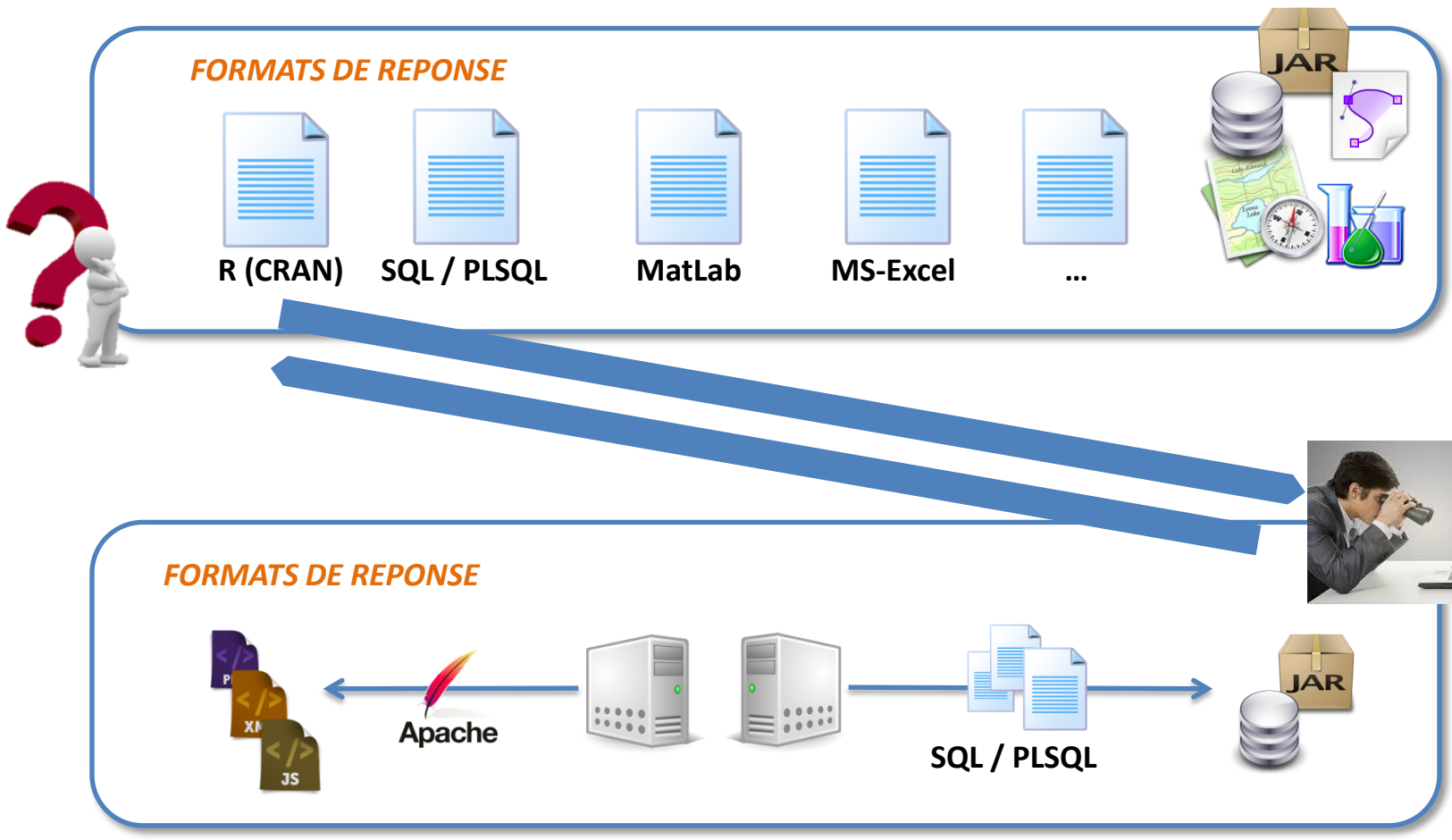
Données manquantes

Saut d'une ou plusieurs années dans l'analyse des séries RCS

Flexibilité et évolutivité de l'OpenSource

… nécessitent des développements et des schémas précis

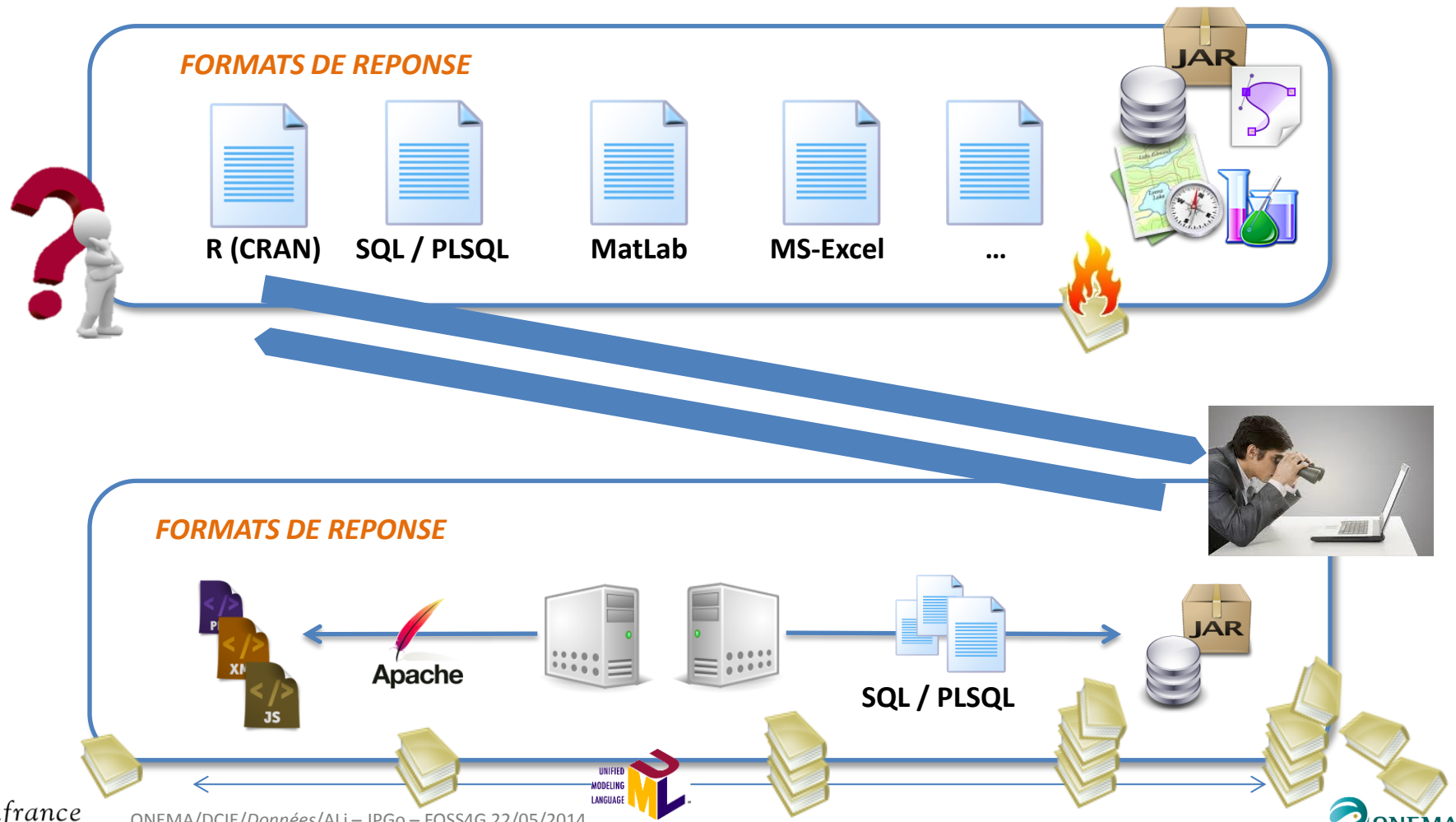
1. Permettre aux ingénieurs d'étude de participer à la résolution des cas



Flexibilité et évolutivité de l'OpenSource

… nécessitent des développements et des schémas précis

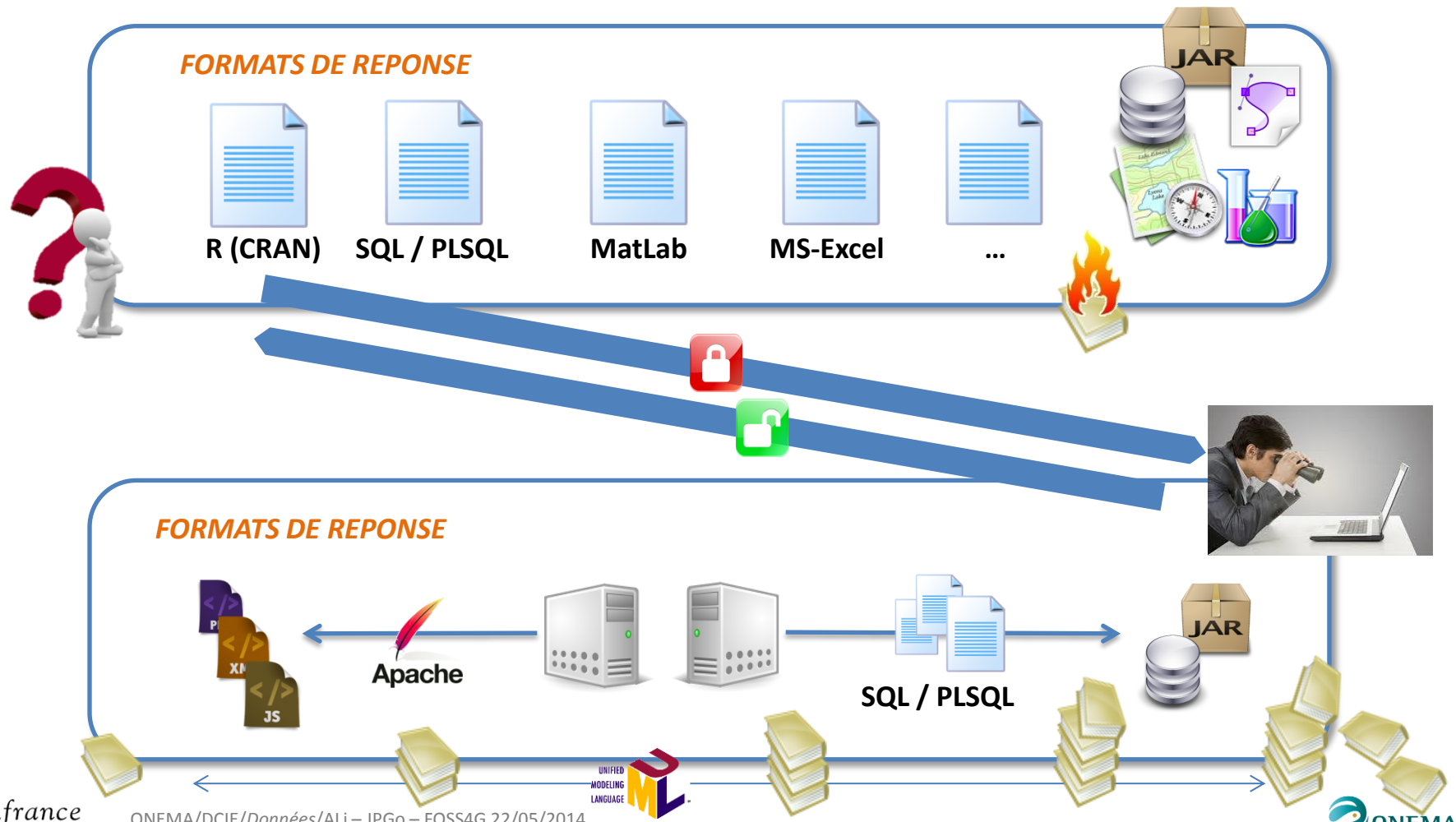
1. Permettre aux ingénieurs d'étude de participer à la résolution des cas



Flexibilité et évolutivité de l'OpenSource

… nécessitent des développements et des schémas précis

1. Permettre aux ingénieurs d'étude de participer à la résolution des cas



Flexibilité et évolutivité de l'OpenSource

… nécessitent des développements et des schémas précis

2. Définir un schéma d'accès aux traitements

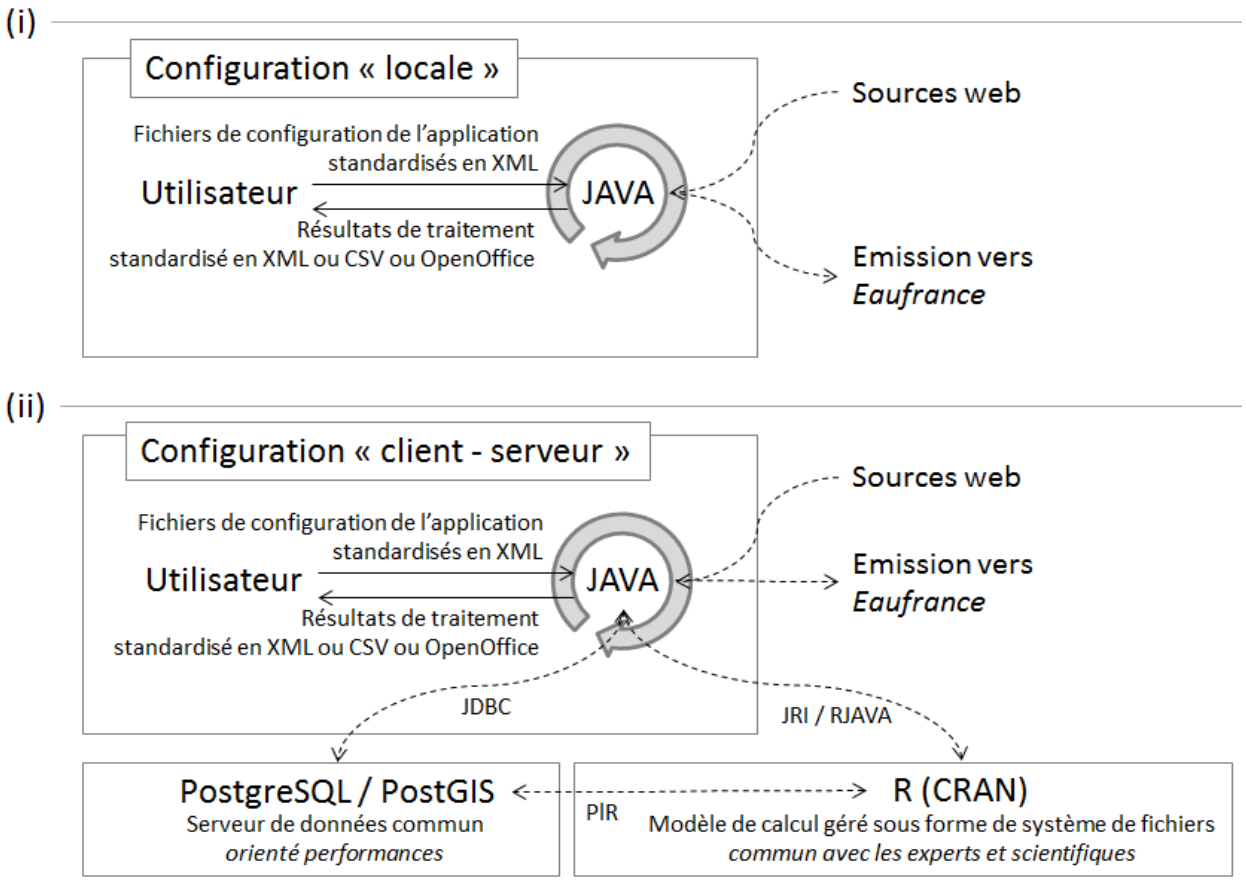


Schéma général proposé pour l'utilisation des technologies Open Source pour le SIE

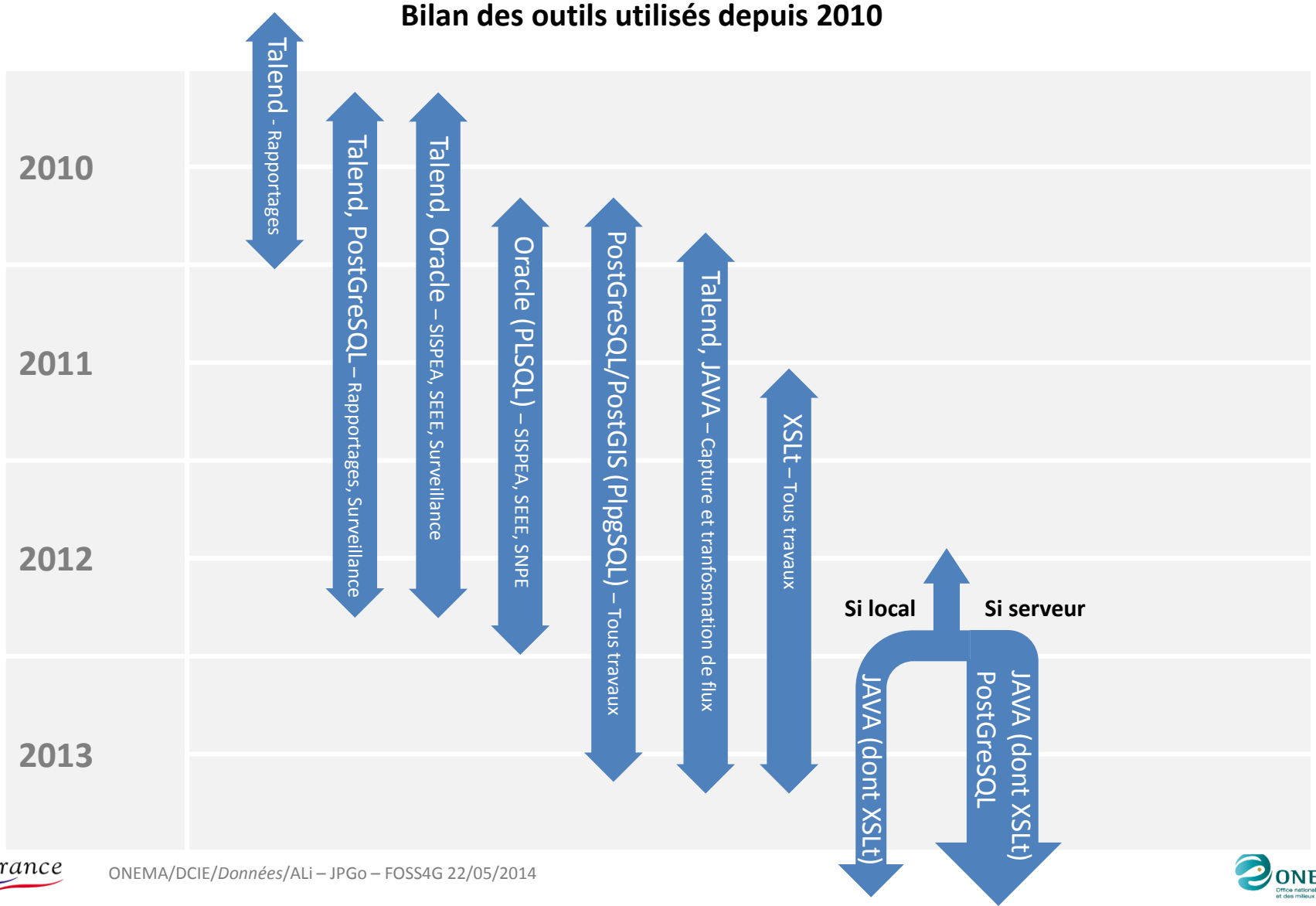
Le cas (i) décrit une utilisation sur poste local, plus limitée car ne mutualisant pas les données, et aux performances dépendant de JAVA (*Apache POI, Xerces* (*SAX*), *DerbyDb, PostgreSQL JDBC driver* et *GeoTools)*.

Le cas (ii) décrit le système en cours de déploiement à la DCIE, utilisant l'interopérabilité des outils de traitements des données.
L'utilisation générique de JAVA en entrée et en sortie de traitement permet une standardisation des traitements, historisés sous forme de fichiers XML, et des produits du système (métadonnées notamment).

PostgreSQL / PostGIS, en pratique

Un choix issu de l'expérience du SIE

Bilan des outils utilisés depuis 2010



Assurer la qualité de l'information

en intégrant les besoins de l'administration de données

Usages

Standards

SANDRE

Quels processus métier conditionnent les données ?

INSPIRE

Formats de référence

ISO19115

Modélisation

Comment pérenniser les données et leur utilisation ?

Contrôle qualité

Métadonnées

Aarhus

ISO19113

Référentiels

Quelle information initiale associée à l'information ?

Quelle information finale associer à l'information ?

Règles juridiques

Directives

Documentation

Une analyse des processus métiers

A la recherche de la cohérence technique et scientifique !

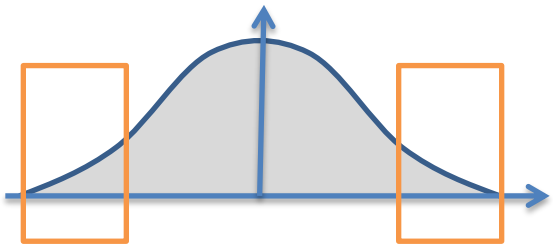
1. Utiliser les statistiques pour qualifier les données

PRINCIPE

✓ choix d'une variable de **distribution**

distribution des stations selon cette variable

modélisation de la loi



identification des cas critiques

✓ **et/ou** implémentation de tests statistiques sur l'ensemble de la population ou des agrégats (Khi-deux, normalité…)

Quelques exemples d'utilisation de statistiques…

Identification d'erreurs d'unités

Séries « chimies » de l'entrepôt du SIE

Identification de valeurs hors bornes, sans avoir d'idée préalable de ces bornes

Description de la relation Taille / Poids des poissons

Représentativité d'une série de données échantillon

Identification et correction de biais pour l'analyse des données de surveillance, de SISPEA…

Une analyse des processus métiers

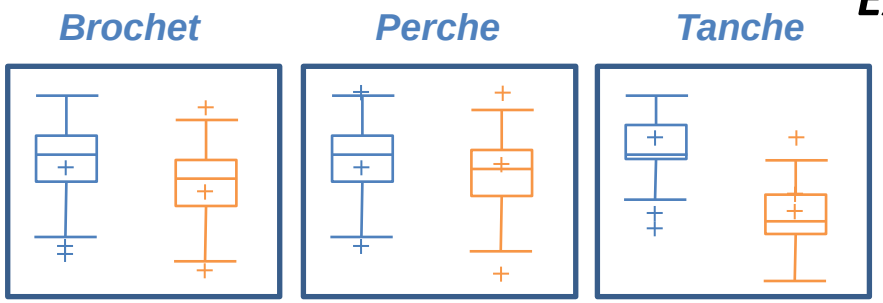
A la recherche de la cohérence technique et scientifique !

Utilise PLR (Joseph E Conway)

1. Utiliser les statistiques pour qualifier les données

EXEMPLE

```
SELECT
        espece_nom,
        r_boxplot(« taille »)
FROM bd_map.series_taille
GROUP BY espece_nom ;
```




```
SELECT
        espece_nom,
        r_boxplot_ident(
                « taille »,
                95)
FROM bd_map.series_taille
GROUP BY espece_nom ;
```

Matrice des points « hors 95 % de la loi normale »

Une analyse des processus métiers

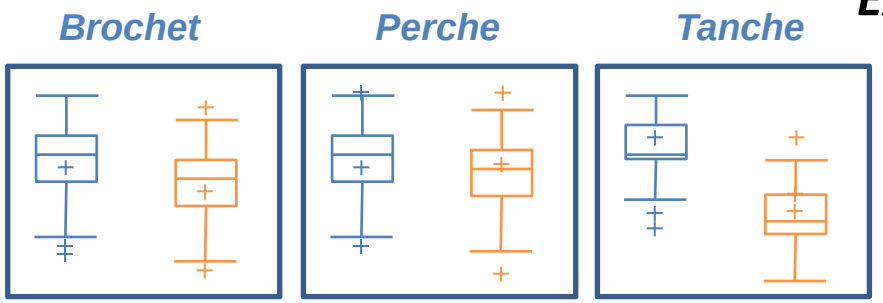
A la recherche de la cohérence technique et scientifique !

Utilise PLR (Joseph E Conway)

1. Utiliser les statistiques pour qualifier les données

EXEMPLE

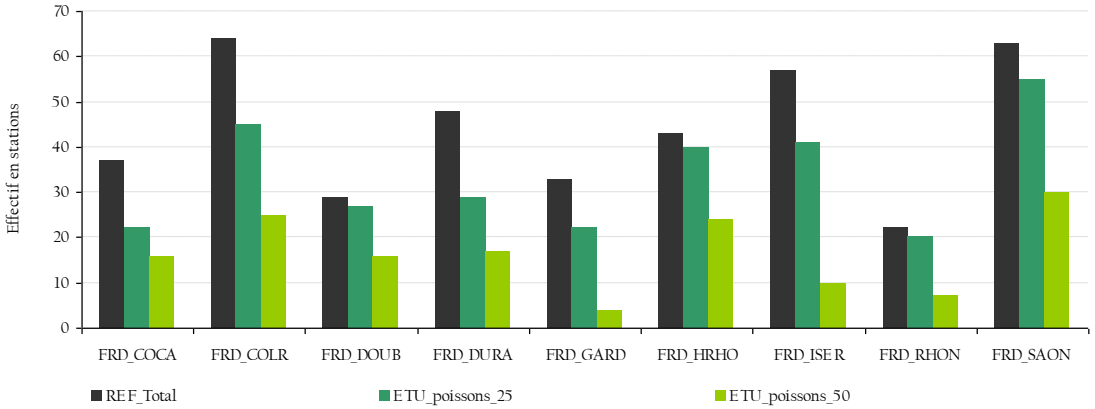
```
SELECT
        espece_nom,
        r_boxplot(« taille »)
FROM bd_map.series_taille
GROUP BY espece_nom ;
```




```
SELECT
        espece_nom,
        r_boxplot_ident(
                « taille »,
                95)
FROM bd_map.series_taille
GROUP BY espece_nom ;
```

Matrice des points « hors 95 % de la loi normale »

```
SELECT
        sous-unite,
        r_khi2(
                « taille »,
                « REF_total »,
                « ETU_poissons_25 »,
                « ETU_poissons_30 »)
FROM surv.stations_pisc
GROUP BY sous-unite;
```



Probabilité que l'échantillon et la série de référence soient issues d'une même population (ici 85 % et 16 %)

Une analyse des processus métiers

A la recherche de la cohérence technique et scientifique !

2. Parcours de réseaux et recomposition de l'information

PRINCIPE

Une partie de l'information est manquante car :

- elle n'a pas été collectée
- un défaut de traçabilité de génération est survenu, et seul le produit final existe

Et on sait qu'en termes d'*information*, le produit disponible permettrait de retrouver ou d'approcher les valeurs initiales.

L'objectif est de reconstituer l'information manquante !

Un exemple précis

Ramener les données « masses d'eau » aux tronçons de la BD Carthage

Les masses d'eau représentent schématiquement un réseau composé de plusieurs tronçons.
Pour ramener les données de la masse d'eau (= groupe de tronçon) à chaque tronçon, il aurait fallu conserver les données d'affectation : ce n'est pas toujours le cas !
Il existe différentes logiques hydrologiques qui permettent de retrouver depuis chaque tronçon, la masse d'eau immédiatement en aval, en parcourant le réseau.

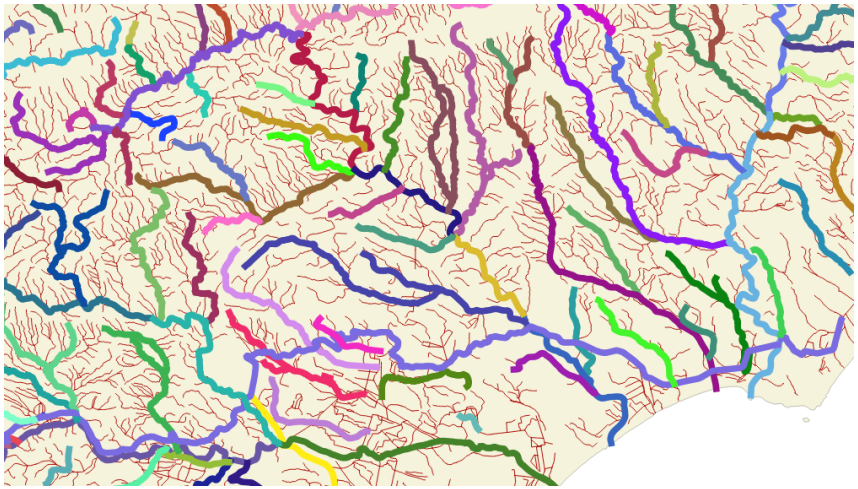
Une analyse des processus métiers

A la recherche de la cohérence technique et scientifique !

Utilise PostGIS 2

2. Parcours de réseaux et recomposition de l'information

EXEMPLE



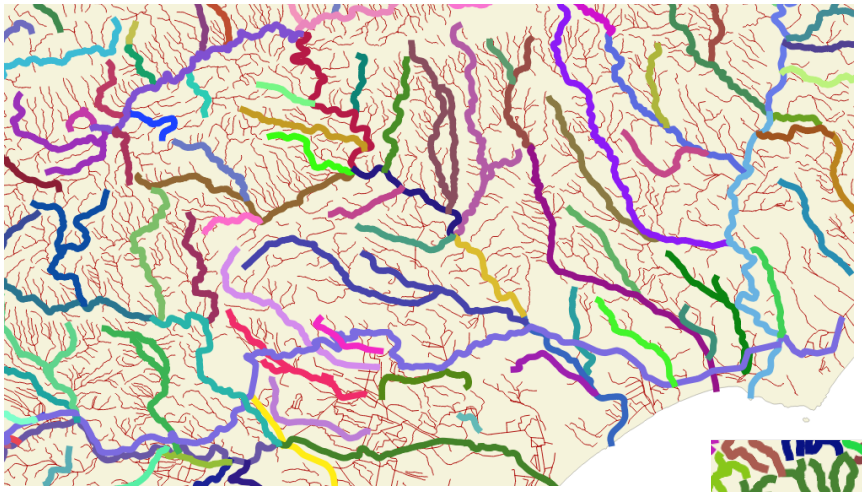
Une analyse des processus métiers

A la recherche de la cohérence technique et scientifique !

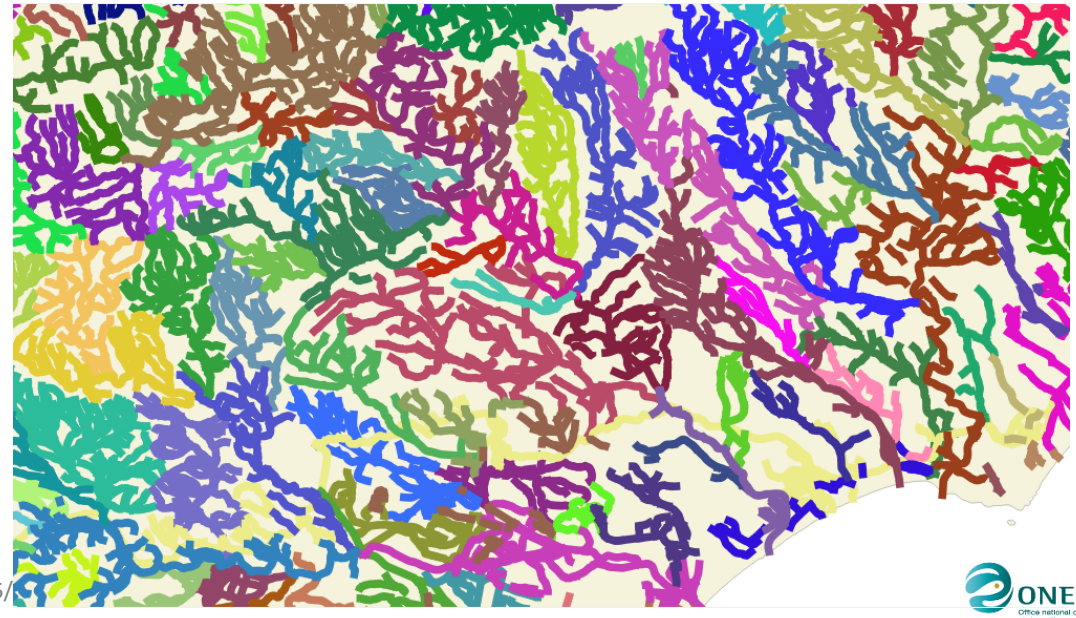
Utilise PostGIS 2

2. Parcours de réseaux et recomposition de l'information

EXEMPLE



Utilisation d'une fonction récursive PlPgSql qui utilise les nœuds pré-renseignés par le BRGM



Compléments d'information par requête IG
Couches fournies par les Agences de l'eau

Vers l'exploitation des données

Interopérabilité des outils : traitements à la volée

Nous avons vu comment PostgreSQL permet l'analyse de la qualité des données, et permet de « retrouver » des informations.

L'aide à la décision et la construction d'indicateurs stratégiques passe par des opérations plus complexes. PostgreSQL permet ici de :

- Générer un **nombre important** d'opérations
- **Standardiser** des approches
- **Tracer** les traitements
- Construire de **nouvelles approches**

→ Intégrer ces approches dans des **applicatifs client-serveur**, une fois automatisés

*Exigence : conserver la possibilité pour l'ingénieur d'étude d'interagir avec le système par un simple dépôt de fichier ! **

* Pas encore opérationnel, attente de PostgreSQL 9.4 pour mobiliser les VARIADIC avec les AGREGATE.

Exploitation statistiques

Faciliter les travaux des ingénieurs

1. Analyses multivariées à la volée

PRINCIPE

Avantages de R :

- des **fonctions statistiques poussées**, issues de la recherche
- une grande souplesse dans les **représentations graphiques**

Une demande récurrente est la réalisation **d'analyses multivariées.**

Les scientifiques mettent au point leurs propres fonctions de représentations. Ils standardisent et génèrent ainsi à la volée une quantité importante de résultats, selon une large gamme de tris.

Un exemple précis

Comparaison d'indicateurs biologiques

Dans l'exemple suivant, on cherche à représenter l'écart entre deux indicateurs proches mais légèrement distincts, afin d'identifier les composantes responsables de leur différence.

Exploitation statistiques

Faciliter les travaux des ingénieurs

Utilise PLR (Joseph E Conway)
Package ade4 CRAN

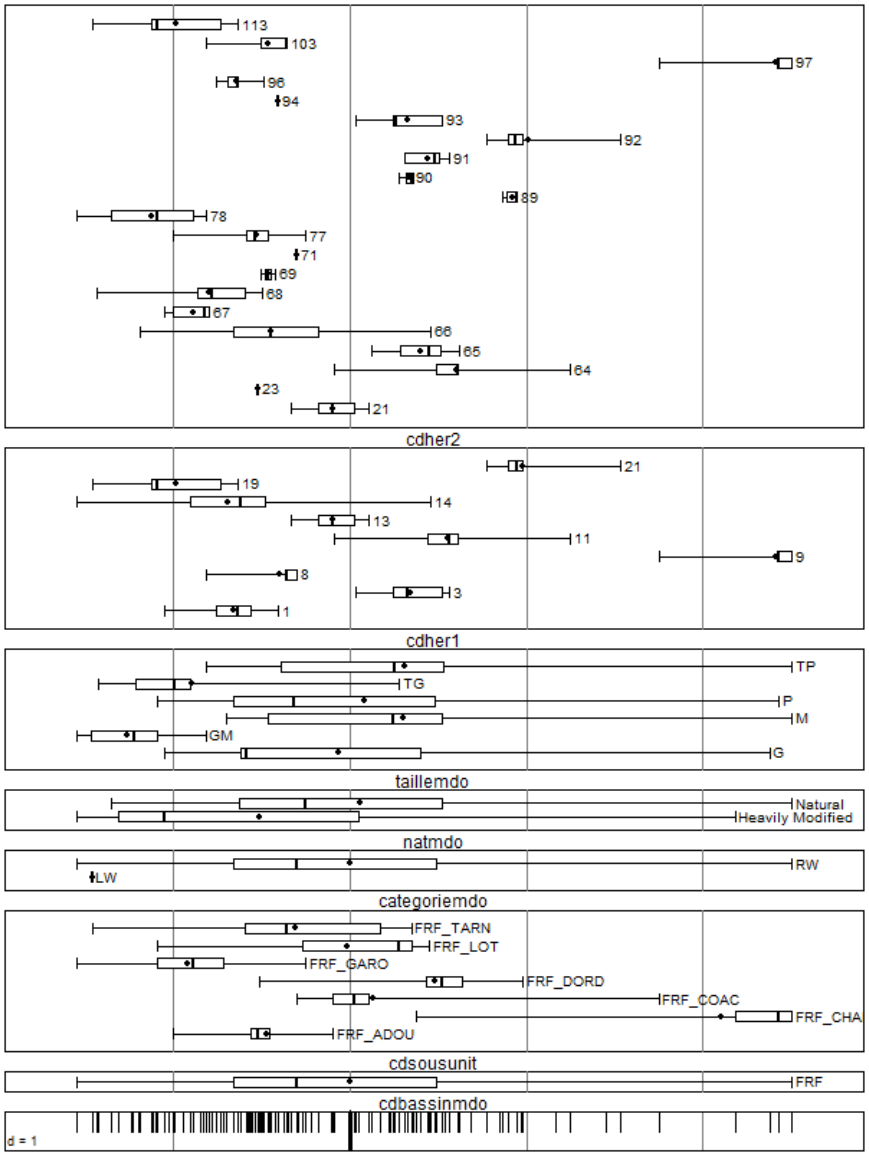
1. Analyses multivariées à la volée

EXEMPLE

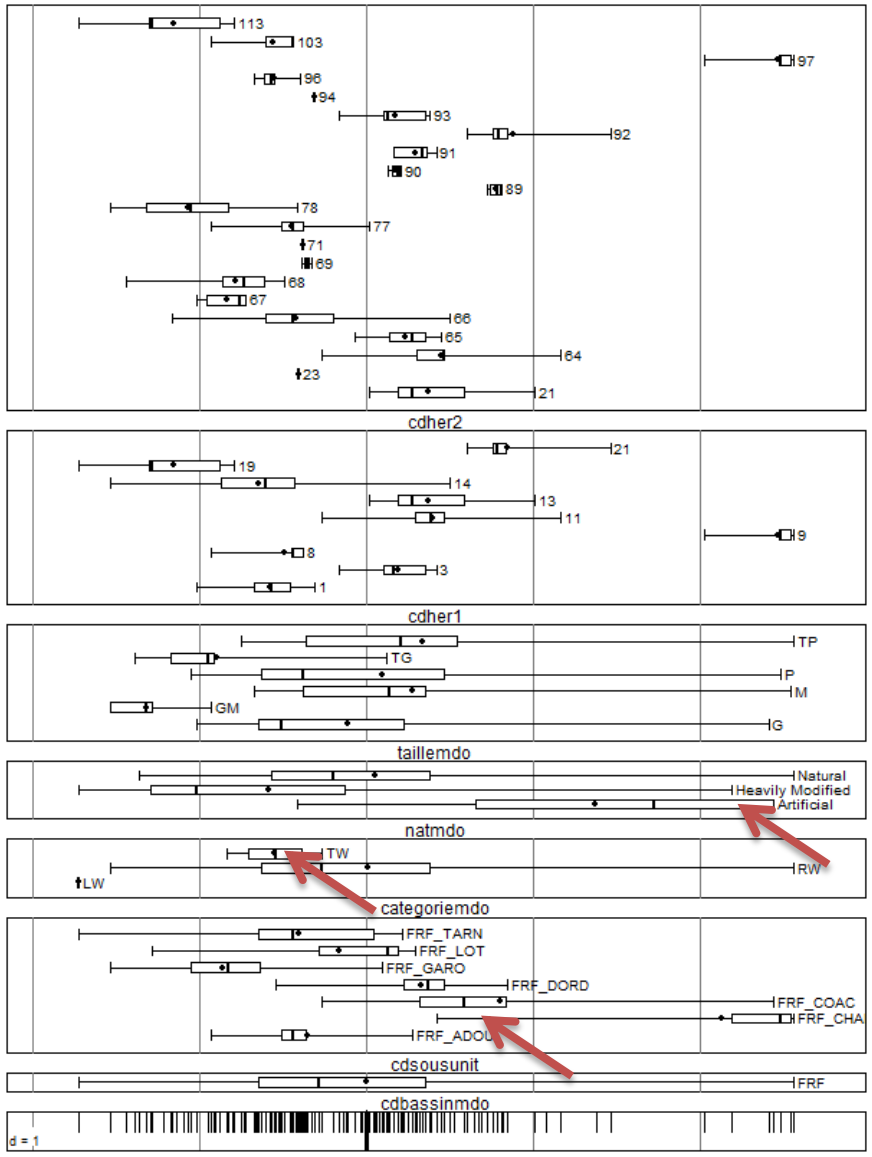
```
SELECT
        r_draw_afc(
                « indice_1 »,
                « cdher2 »,
                « cdher1 »,
                « tailemdo »,
                « natmdo »,
                « categoriemdo »,
                « cdsousunit »,
                « cdbassinmdo »,
                ‘spearman’
        )
FROM  surv.serie_indice1;
```

```
SELECT
        r_draw_afc(
                « indice_2 »,
                « cdher2 »,
                « cdher1 »,
                « tailemdo »,
                « natmdo »,
                « categoriemdo »,
                « cdsousunit »,
                « cdbassinmdo »,
                ‘spearman’
        )
FROM  surv.serie_indice2;
```

Indicateur biologique 1



Indicateur biologique 2



2. Génération de cartes statistiques *à la volée*

PRINCIPE

Les utilisateurs peuvent avoir besoin de projeter les **informations issues de calculs sur des fonds de carte de référence**.

- *Ces calculs peuvent être issus du code SQL, ou de R (exemple : médiane)*

- *R peut interpréter les formats WKT et produire du SVG, afin d'utiliser des fonctions d'interpolation puissantes embarquées par les bibliothèques d'interpolation (ce qui évite les transferts supplémentaires de données)*

Un exemple précis

Représentations cartographique de l'Indice Poisson Rivière

On souhaite disposer de la médiane par zone hydrographique d'une part, d'un modèle numérique de terrain d'autre part.

Exploitation statistiques

Faciliter les travaux des ingénieurs

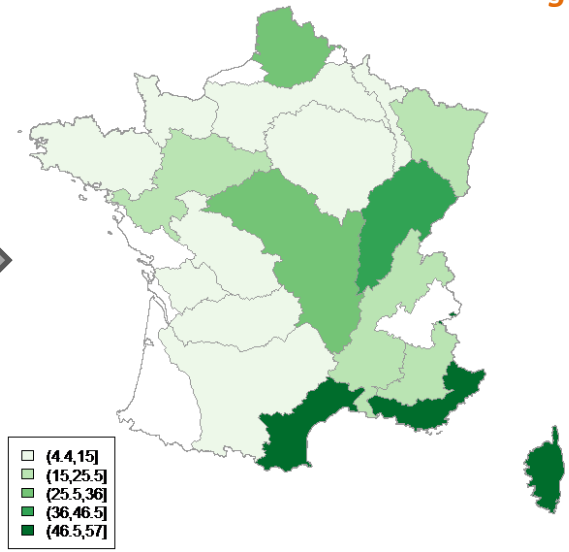
Utilise PLR (Joseph E Conway)
Packages sp, rgeos, maptools, gstat, RColorBrewer CRAN

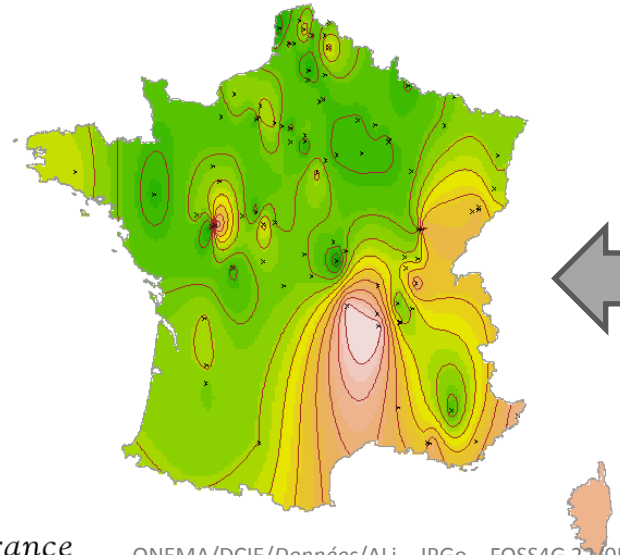
EXEMPLE

Mobilise du WKT, produit du SVG

2. Génération de cartes statistiques *à la volée*

```
SELECT
        region_hydro,
        r_draw_statmap (
                « ipr »,
                'median',
                ST_AsText(« geom »)
        )
FROM  surv.serie_ipr
GROUP BY region_hydro;
```





```
SELECT
        r_draw_interpolmap (
                « ipr »,
                'linear',
                ST_AsText(« geom »)
        )
FROM  surv.serie_ip;
```

Ces cartes se redimensionnent si la requête d'entrée est limitée par une condition « WHERE ».

En pratique, R crée un fichier sur le disque et retourne l'adresse à PostgreSQL.

L'équipe projet
Données pour la décision

alexandre.liccardi@onema.fr
jean-philippe.goyen@onema.fr

Laurent Coudercy
Chef de département
Directeur du Pôle INSIDE

Alexandre Liccardi
Chef de projet

Jean-Philippe Goyen
Chargé d'études *Données du SIE*

Participants

Laurent Breton
Chef de projet IG

Jérôme Bouche
Chargé d'étude R (CRAN)

Merci de votre attention