

# INTRODUCTION AUX BASES DE DONNÉES EN BIODIVERSITÉ

Atelier CEBioS MRV, Septembre 2015

---

André Heughebaert  
[Belgian Biodiversity Platform](#)

# CONTENU

---

1. Introduction
2. Données
3. Metadonnées
4. SQL
5. NoSQL
6. Outils
7. Standards

# 1. INTRODUCTION



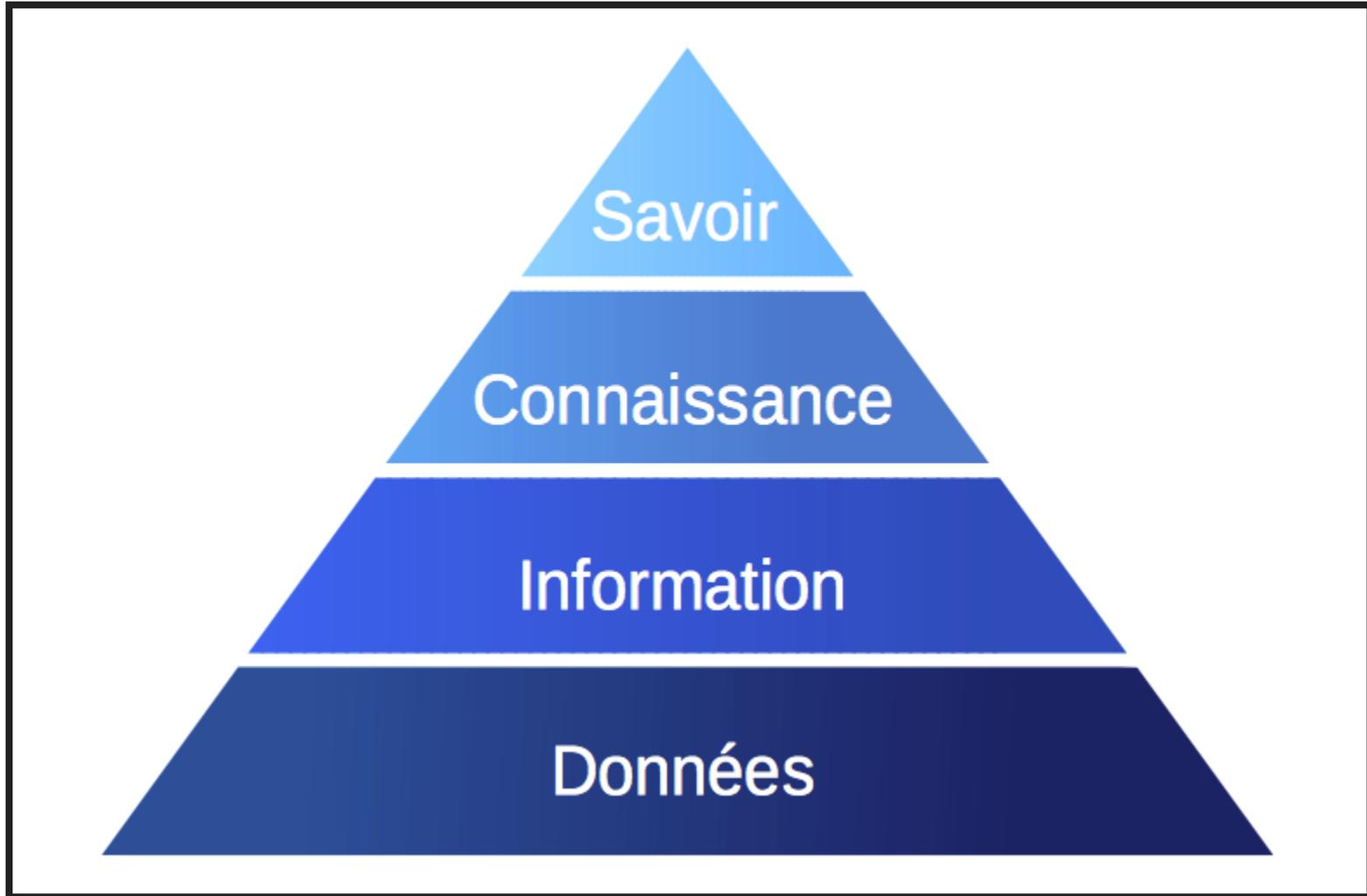
# DONNÉES EN BIODIVERSITÉ

---

- Taxonomie, noms vernaculaires, specimens types
  - Population, habitat, aire de répartition
  - Description, traits, propriétés, génome
  - Relation entre espèces, écosystèmes
  - Observations, collections
  - Indicateurs, modèles
- 
- Publications scientifiques (et autres)
  - Multimedia: images, sons, vidéos...
  - Droit: Législations, réglementations, aires protégées...
  - Communauté d'experts, scientifiques, amateurs

# DONNÉES-INFORMATION

---



# DONNÉES-INFORMATION

---

Une **donnée** est une description élémentaire d'une réalité.

C'est par exemple une observation ou une mesure.

La donnée est dépourvue de tout raisonnement, supposition, constatation, probabilité.

L'**information** est une donnée compréhensible, qui a un sens pour la personne qui la recoit.

La **connaissance** est le résultat d'une réflexion sur les informations analysées.

Le **savoir** permet le discernement final sur le contenu (informations et connaissances) et le jugement de bon sens.

# DONNÉES-MÉTADONNÉES

---

**Les données (primaires) : Quoi?, Quand?, Où?, Combien?...**

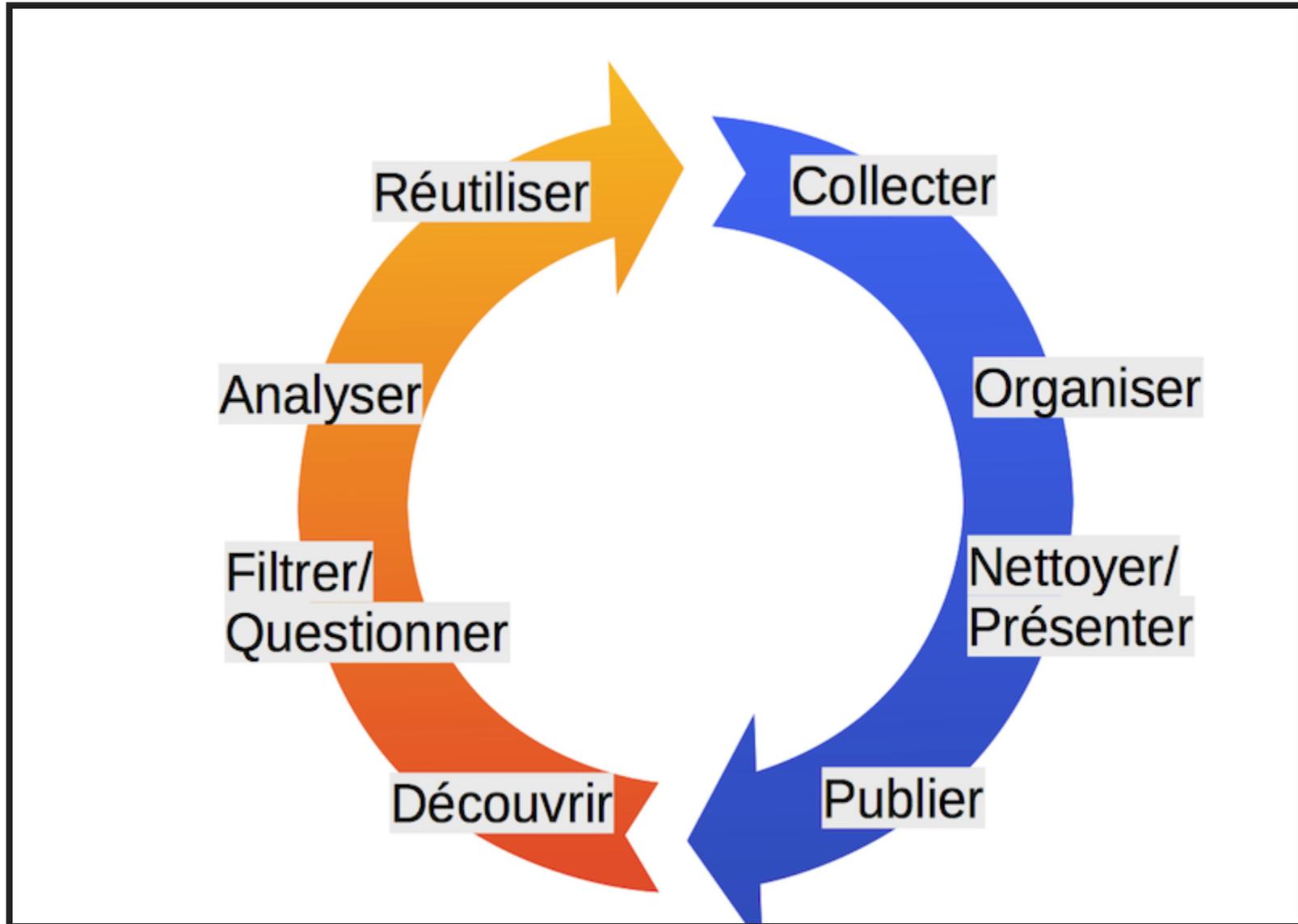
**Les métadonnées : Qui?, Comment? Pourquoi?...**

---

*Pas de données sans métadonnées!*

# CYCLE DE VIE DES DONNÉES

---



## 2. DONNÉES (NOTIONS DE BASE)

---

- Espèces, Individus, Habitats, Ecosystèmes, Menaces, Indicateurs... sont des **entités** différentes
- Chaque entité a des **attributs**: par exemple le nom, l'aire de répartition ou le status de protection pour une espèce.

## 2. DONNÉES (NOTIONS DE BASE)

---

- Il existe des **relations entre entités**
  - telle espèce vit dans tel habitat
  - tel individu appartient à telle espèce
  - telle espèce parasite telle autre espèce
- Il existe des **contraintes sur les attributs**
  - un nom scientifique est exprimé par un binôme latin, il est unique!
  - une latitude est comprise entre  $-90^\circ$  et  $+90^\circ$ .
  - une longitude est comprise entre  $-180^\circ$  et  $+180^\circ$ .

### 3. METADONNÉES

---

Metadonnées: **données descriptives** d'un jeu de données.

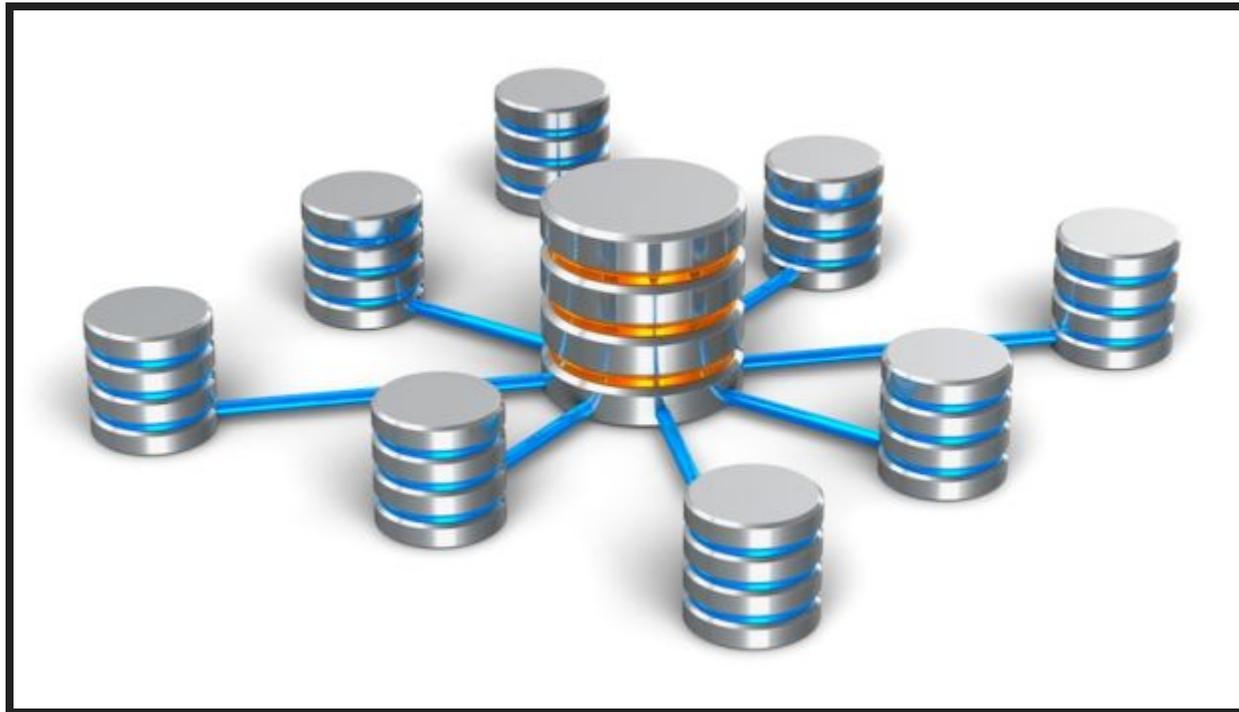
- **Qui?**: nom, prénom, institution, partenaires
- **Comment?**: méthodologie, matériel, fréquence, limites...
- **Pourquoi?**: projet de recherche ou de monitoring, source de financement...

### 3. METADONNÉES

1. Métadonnées de base
  2. Géographiques
  3. Temporels
  4. Taxonomiques
  5. Mots-clés
  6. Description du projet
  7. Méthode d'échantillonnage
-

## 4. SQL (BASE DE DONNÉES RELATIONNELLE)

---

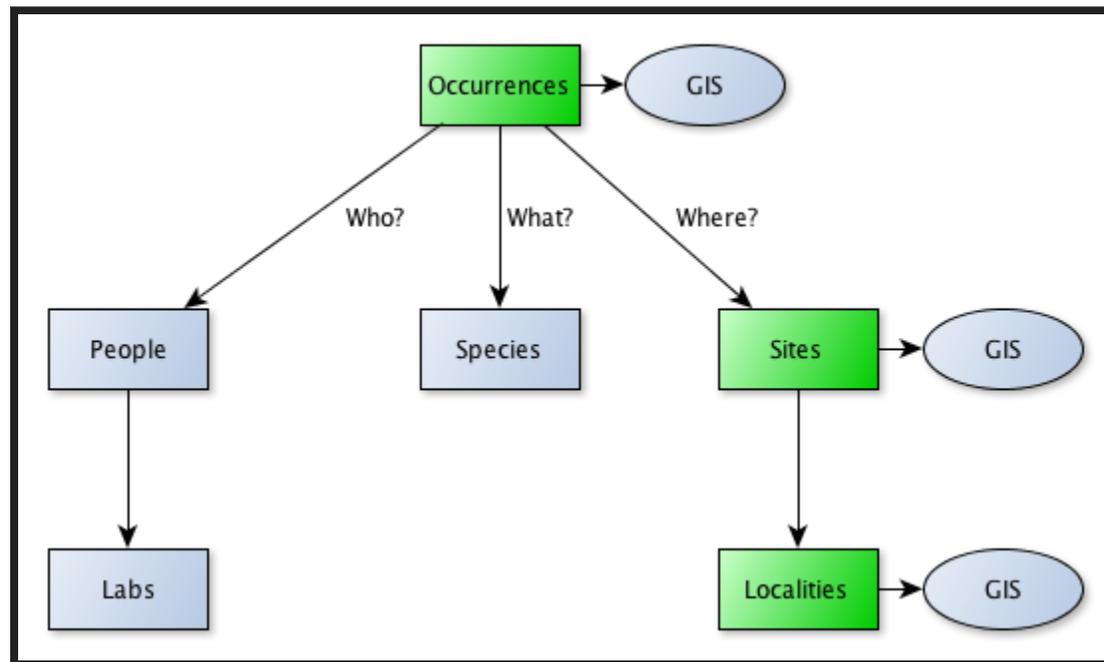


*Organisez vos données suivants vos  
besoins.*

---

# 4. SQL (SCHÉMA DE DONNÉES)

---



## 4. SQL (NOTIONS DE BASE)

---

Entité

Table

---

Attribut

Colonne

---

Relation

Clé+index

---

Contrainte

Contrainte

## 4. SQL (NOTIONS DE BASE)

---

- SQL=Structured Query Language, en français langage de requête structurée.
- SQL permet de rechercher, d'ajouter, de modifier ou de supprimer des données.
- La base de données vérifie automatiquement toute règle ou contrainte et garantit ainsi **l'intégrité de vos données.**

## 4. SQL (CREATE)

```
create table people (  
    peo_id integer primary key,  
    lab_id integer references labs,  
    peo_familyName text,  
    peo_firstName text,  
    peo_beginDate text,  
    peo_endDate text);  
  
create table sites (  
    sit_id integer primary key,  
    loc_id integer references localities,  
    sit_name text,  
    sit_latitude real,  
    sit_longitude real,  
    sit_uncertainty integer);
```

## 4. SQL (SELECT)

---

Les **requêtes** permettent d'interroger la base de données.

```
select * from occurrences where occ_date > '1830-01-01';  
  
select occ.id, occ_latitude, occ_longitude from occurrences  
where occ_latitude is not null;
```

## 4. SQL (JOIN)

---

Les **jointures** permettent d'associer plusieurs tables dans une même requête.

```
select * from occurrences occ
left join people peo on peo.peo_id = occ.peo_id
where (occ.occ_date < peo.peo_beginDate) or
(occ.occ_date > peo.peo_endDate);
```

## 4. SQL (VIEW)

---

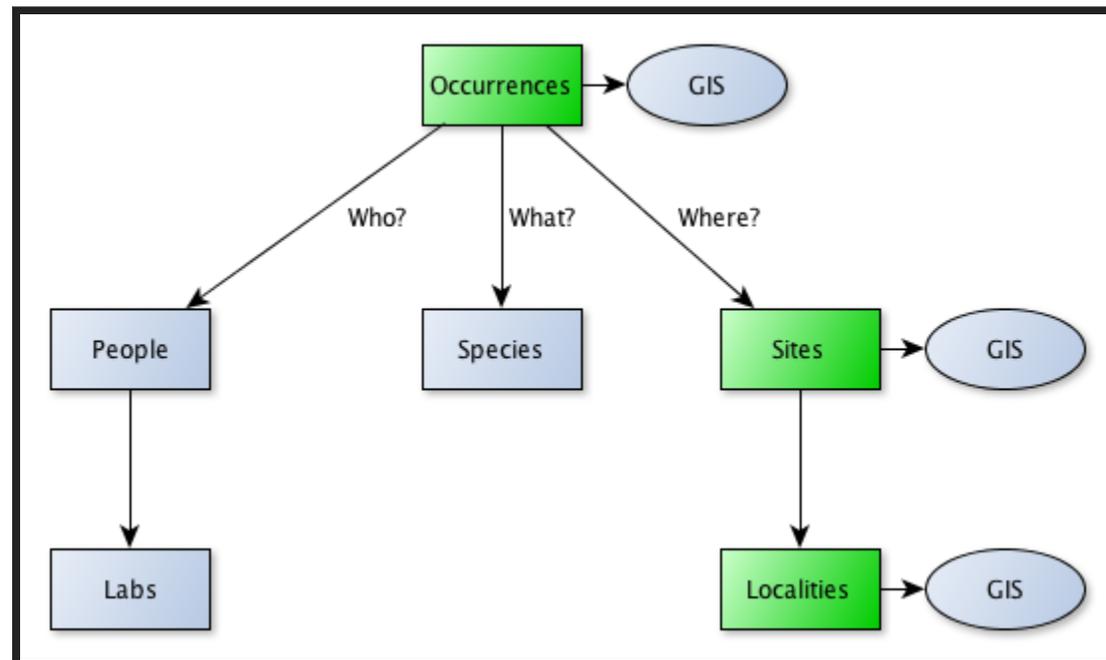
Les **vues** présentent les résultats d'une requête sous forme d'une table virtuelle.

```
CREATE VIEW observations AS
select occ.occ_id AS id, spe.spe_scientificName AS scientificName, spe
coalesce(occ.occ_uncertainty,sit.sit_uncertainty) AS uncertainty, loc
from (((occurrences occ left join sites sit on((sit.sit_id = occ.si
left join localities loc on((loc.loc_id = sit.loc_id)))
left join species spe on((spe.spe_id = occ.spe_id)))
left join people peo on((peo.peo_id = occ.peo_id)))
left join labs lab on((lab.lab_id = peo.lab_id)));

select * from observations;
```

# 4. SQL (DEMO)

---



## 4. SQL (RÈGLES D'OR)

*Gardez vos entités(=tables) simples.*

*Une seule information par colonne.*

*Utilisez des **identifiants uniques** et  
**persistants**.*

*Utilisez des noms simples et parlants.*

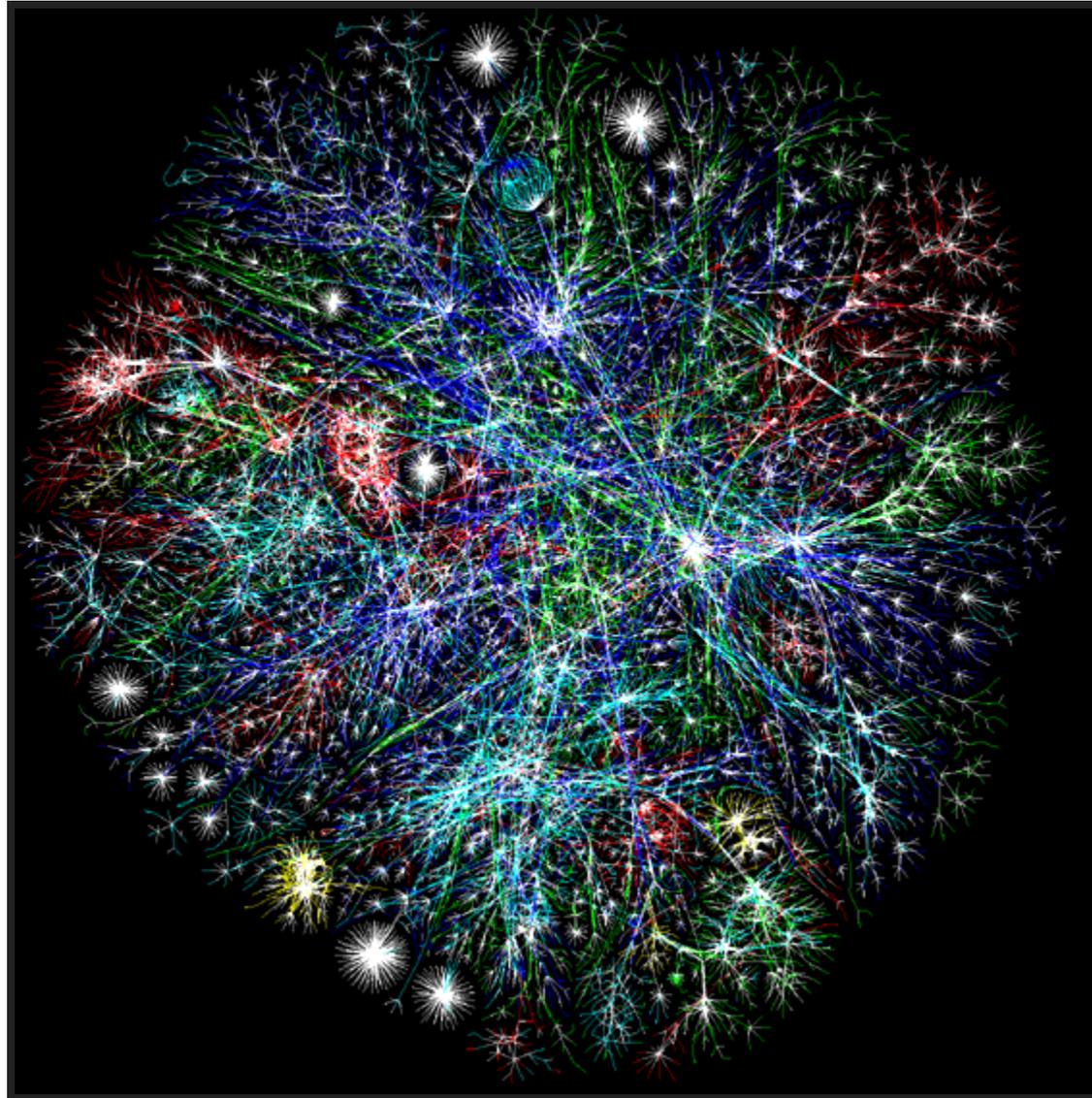
*Evitez les redondances, source d'erreur.*

*Utilisez des **vocabulaires fermés**  
(eg code ISO 3166-1-alpha-2 pour pays).*

---

# 5. NOSQL

---



## 5. NOSQL

A côté des base de données relationnelles, ils existent d'autres solutions pour organiser/interroger vos données:

- Clés-valeurs
- Triplets, RDF
- Graph database
- Base de données orientées objets
- ...

## 6. OUTILS

---



---

*Être à l'aise avec vos outils va grandement vous simplifier la vie.*

## 6. OUTILS (RECOMMANDÉS)

- Editeur de texte (pour vos données et vos scripts)
- Tableur: [LibreOffice](#), [OpenRefine](#)
- Base de données: [SQLite](#), [PostgreSQL](#)
- Cartographie: [QGIS](#)
- (Language: Java, Python, Ruby...)



## 6. OUTILS

---



## 7. STANDARDS (D'ÉCHANGES DE DONNÉES)

- **JSON** est un format texte léger; facile à lire ou à écrire pour des humains et aisément analysable ou générable par des machines.
- **CSV**(ou TSV) format texte de données tabulaires. Chaque ligne du texte correspond à une ligne du tableau et les virgules(ou tab) correspondent aux séparations entre les colonnes.
- Darwin Core (**DwC**) est un ensemble de standards pour le partage des données de biodiversité. Il a été développé par le groupe international TDWG.

# DES QUESTIONS?

Belgian Biodiversity Platform

[a.heughebaert@biodiversity.be](mailto:a.heughebaert@biodiversity.be)



Fait avec [reveal.js](https://revealjs.com/) - HTML Presentations made easy