Pôle national des données de biodiversité

Yvan Le Bras
yvan.le-bras@mnhn.fr

@Yvan2935
Station de Biologie marine Concarneau

<>)))><
—

UMS 2006 PatriNat MNHN/CNRS/AFB
Jardin des plantes 75005 Paris

« Datanami » en sciences

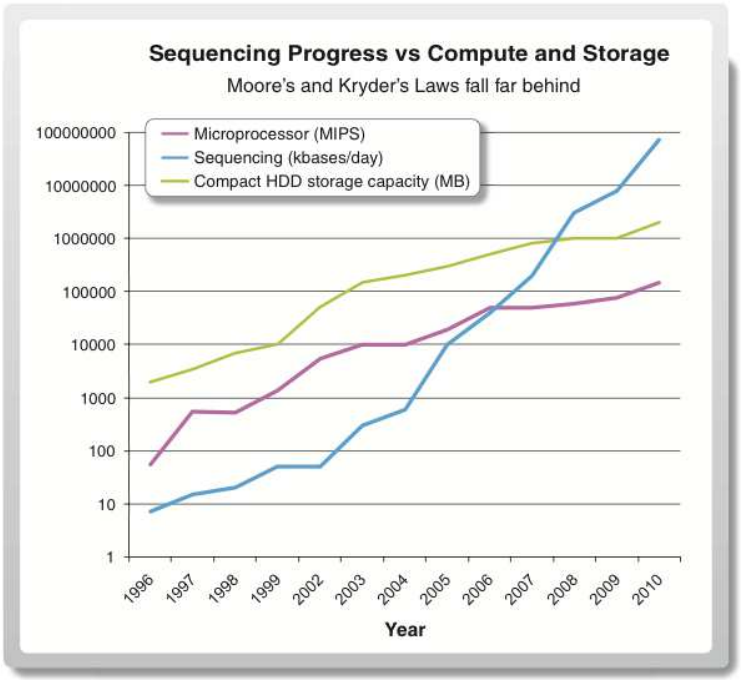


Fig. 1. A doubling of sequencing output every 9 months has outpaced and overtaken performance improvements within the disk storage and high-performance computation fields.

Kahn. On the future of genomic data. Science (2011) vol. 331 (6018) pp. 728-9

- *omique : Séquençage nouvelle génération / Protéomique / Métabolomique
- Ecologie
- Donnée digitale
  - Quantité
  - Hétérogénéité
- Situation critique pour les laboratoires
- La solution = optimiser et mutualiser !

PNDB : Fonction Services et outils

Données et métadonnées ***« as open as possible, as closed as necessary »***

- Accès aux données primaires brutes, aussi ouvertes que possibles et privilégier données ouvertes (pour use case notamment)
- Choix d'un standard unique pivot, l'EML
- *Identification des ressources terminologiques et ontologiques existantes et pertinentes*
- *Mise à disposition simplifiée des ressources terminologiques et ontologiques existantes et pertinentes*
- *Outil de mapping automatique/semi-automatique de métadonnées en lien avec contenu des données*
- *Mise en place système permettant / facilitant le retour d'utilisateurs et/ou équipe PNDB vers les contributeurs (scientifiques/IR/Partenaires)*

Traitement ***« as easy as possible, as complicated as necessary »***

- Accès plateforme Galaxy-E
- *Investir dans le développement de nouveaux outils / nouvelles fonctionnalités Galaxy-E*
- *Mise à disposition outils de traitements « classique » mais via web (R/RStudio/R Shiny/Jupyter Notebook)*

Orientation FAIR

- Application des principes FAIR (FORCE11)
  - **F**

TO BE FINDABLE:

F1. (meta)data are assigned a globally unique and eternally persistent identifier.
F2. data are described with rich metadata.
F3. (meta)data are registered or indexed in a searchable resource.
F4. metadata specify the data identifier.

  - **A**

TO BE ACCESSIBLE:

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
A1.1 the protocol is open, free, and universally implementable.
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
A2 metadata are accessible, even when the data are no longer available.

  - **I**

TO BE INTEROPERABLE:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles.
I3. (meta)data include qualified references to other (meta)data.

  - **R**

TO BE RE-USABLE:

R1. meta(data) have a plurality of accurate and relevant attributes.
R1.1. (meta)data are released with a clear and accessible data usage license.
R1.2. (meta)data are associated with their provenance.
R1.3. (meta)data meet domain-relevant community standards.

Orientation FAIR

- Application des principes FAIR (FORCE11)
  - RDA FAIR Data Action Plan : https://zenodo.org/record/1285290#.W59qi_Y69hE

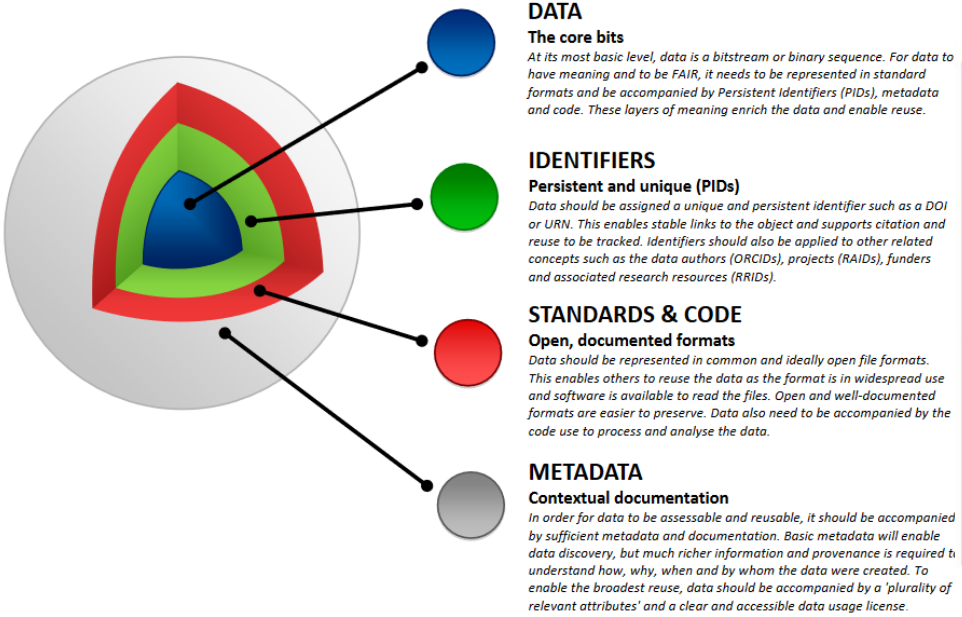


Figure 1: A model for FAIR Data Objects, noting the elements that need to be in place for data to be Findable, Accessible, Interoperable and Reusable.

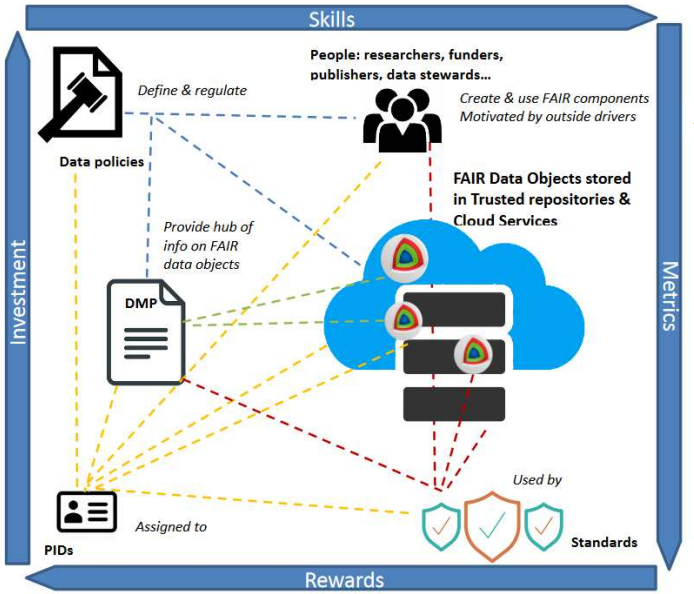


Figure 3: The interactions between components in the FAIR data ecosystem. Registries need to sit behind each component to support automated workflows across them.

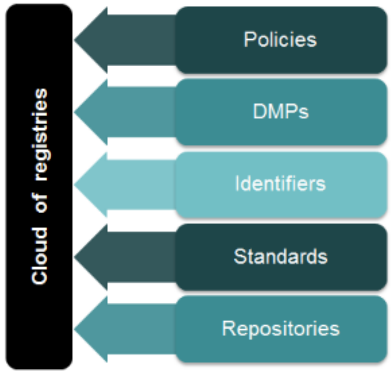


Figure 2: The components of a FAIR data ecosystem

Step 4: Embed a culture of FAIR in research practice

Rec. 10: Trusted Digital Repositories

Repositories need to be encouraged and supported to achieve CoreTrustSeal certification. The development of rival repository accreditation schemes, based solely on the FAIR principles, should be discouraged.

Rec. 12: Data Management via DMPs

Any research project should include data management as a core element necessary for the delivery of its scientific objectives, addressing this in a Data Management Plan. The DMP should be regularly updated to provide a hub of information on the FAIR data objects.

Rec. 16: Broad application of FAIR

FAIR should be applied broadly to all objects (including metadata, identifiers, software and DMPs) that are essential to the practice of research, and should inform metrics relating directly to these objects.

Rec. 19: Encourage and incentivise data reuse

Funders should incentivise data reuse by promoting this in funding calls and requiring research communities to seek and build on existing data wherever possible.

Principale difficulté

- **Associer des métadonnées structurées et détaillées aux données de biodiversité**

Data

★★★ Data accessible on the Web, with an open license, in an open format

Metadata

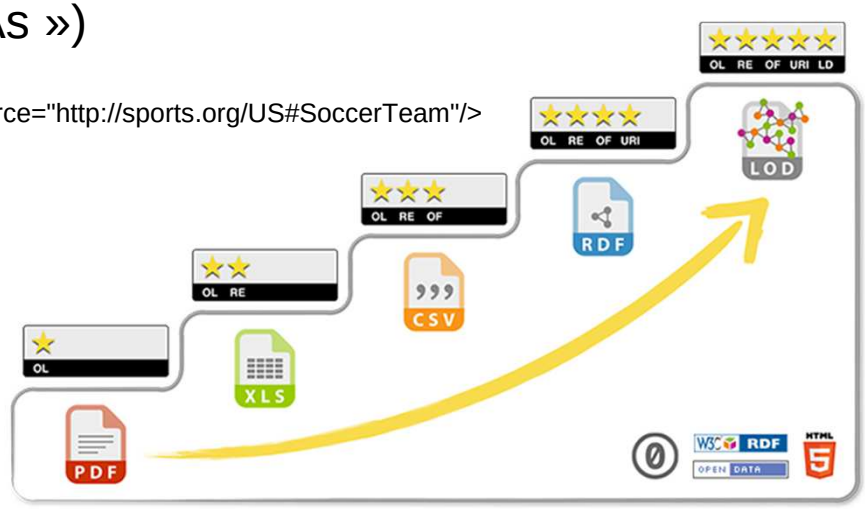
★★★★★ Data **DESCRIBED** and **LINKED** with RDF

- Open Data
  - In France & science: CC-BY 4.0 or Etalab
- Linked data
  - external references through URI / DOI
  - use of « owl:sameAs »)

```
<owl:Class rdf:ID="FootballTeam">
        <owl:sameAs rdf:resource="http://sports.org/US#SoccerTeam"/>
</owl:Class>
```

Data

Metadata



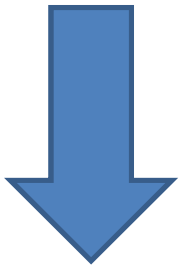
Principale difficulté

- **Associer des métadonnées structurées et détaillées aux données de biodiversité**

Data

Metadata

- « Complete » description
  - Sufficient for Re-analysis
- Based on / extensive use of controlled vocabularies
  - Ontologies
  - Thesaurus

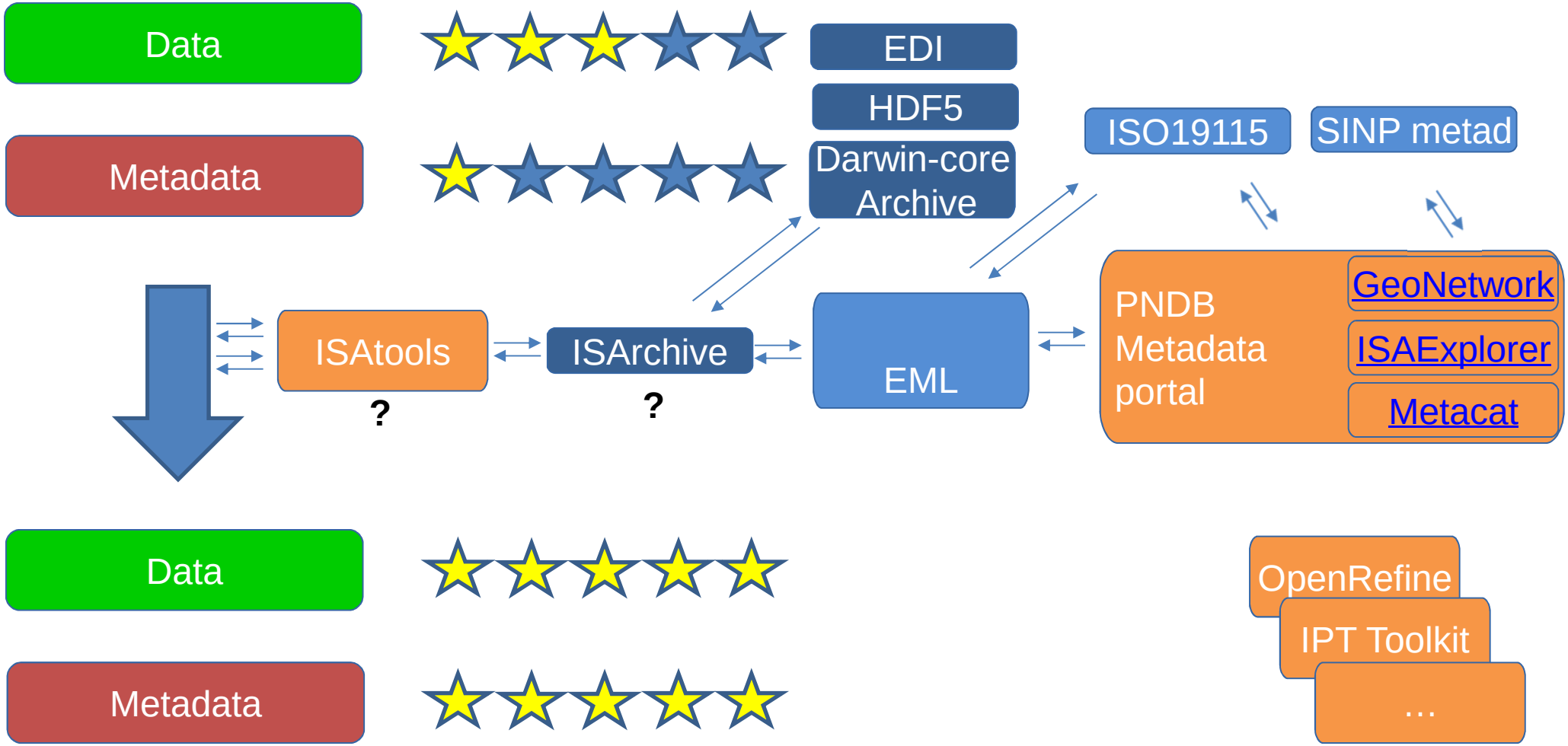


Data

Metadata

Proposition PNDB « data & metadata »

- **Associer des métadonnées structurées et détaillées aux données de biodiversité**



Proposition PNDB « data & metadata »

- **Association données/métadonnées via data package?**