

# Big Data et IA dans la recherche

## Enjeux épistémologiques

---

Charles H. Pence •  @pence@scholar.social

Colloque du CEP, 2025-04-10



# Plan

1. Le *Big Data* dans les sciences
2. Quelle est la nature de ces données ?
3. L'analyse des données
  - 3.1 Problèmes avec les données
  - 3.2 Problèmes avec l'éthique
  - 3.3 Problèmes avec la connaissance

**Le message :** En plus de l'éthique, l'IA change aussi l'épistémologie là où elle est déployée.

# Le *Big Data* dans les sciences

---

In October 2018, the CERN Data Centre passed the milestone of 300 petabytes of data permanently archived in its tape libraries. At the end of 2018, 330 PB of data were permanently archived on tapes in the CERN Data Centre.

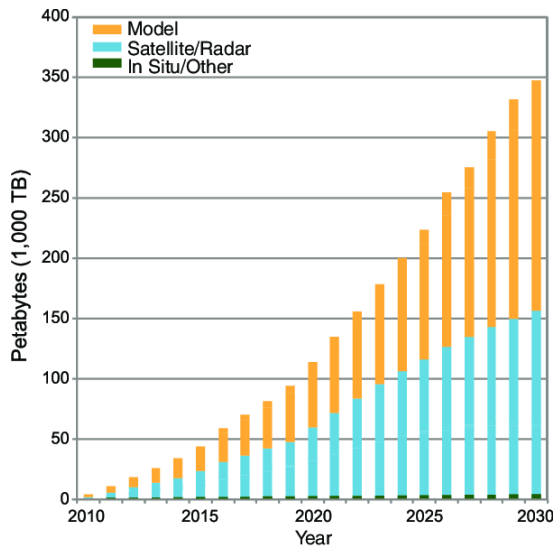
New record in November 2018 for data taking over a single month: 15.8 petabytes of data (from all sources) were written on tape that month.

Within one year, when the LHC is running, more than one exabyte (the equivalent to 1000 petabytes) of data is being accessed (read or written).

gigabyte → terabyte (1000 GiB) →  
petabyte (1000 TiB) → exabyte (1000 PiB)

Estimates predict that genomics research will generate between 2 and 40 exabytes of data within the next decade. (NHGRI)

# En climatologie



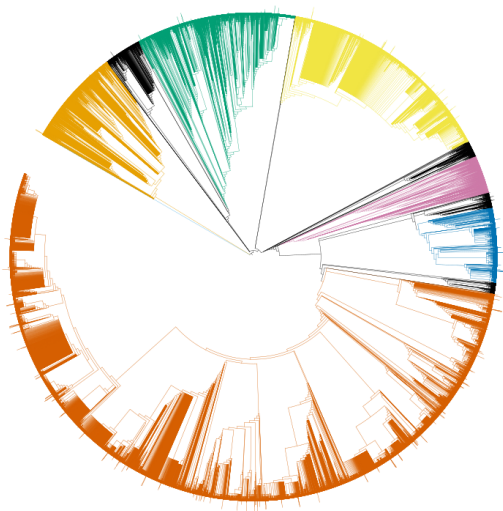
# Qu'est-ce qu'une donnée ?

---

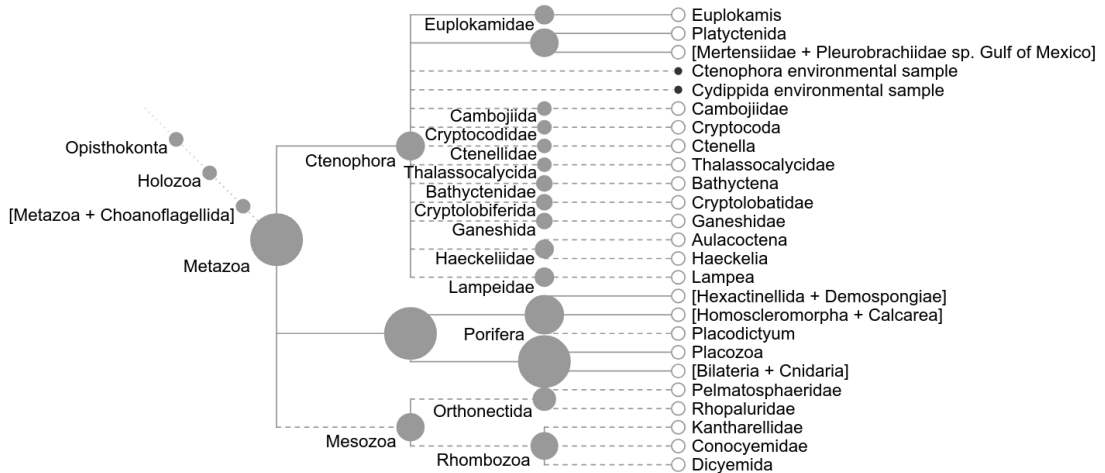
# Des « données » nouvelles

- De questions de « création » aux questions d'« accès » et d'« attribution »
- La mort de **l'obscurité pratique** et de **l'anonymat automatique**
- Surveillance en continu et normalisée
- Données qui savent plus que nous

# Cas d'étude : The Open Tree of Life



# Cas d'étude : The Open Tree of Life



# Cas d'étude : The Open Tree of Life

**Bathyctena chuni** ×

**Node properties**

**Source taxonomy**

[NCBI: 1403704](#)  
[WoRMS: 265844](#)  
[GBIF: 2501150](#)  
[IRMNG: 11329585](#)

**Reference taxonomy**

[OTT: 3684140](#)

**Node id in synthetic tree**

[ott3684140](#)

**Leaf node (no descendant tips)**

[Download subtree as Newick string](#)

[Search EOL for 'Bathyctena chuni'](#)

[Browse Bathyctena chuni in OneZoom](#)

**Edge to parent *Bathyctena***

**Supporting trees** [\[show all details\]](#)

Taxonomy [ott3.7draft2](#)

# Cas d'étude : The Open Tree of Life

NCBI Taxonomy Browser

Search for:  as: complete name ☒ lock

Display: 3 levels using filter: none

**Bathytectena chuni**

Taxonomy ID: 1403704 (for references in articles please use NCBI:txid1403704)

current name

*Bathytectena chuni* (Moser, 1909)  
basonym: *Mertensia chuni* Moser, 1909

NCBI BLAST name: **comb jellies**

Rank: **species**

Genetic code: [Translation table 1 \(Standard\)](#)

Mitochondrial genetic code: [Translation table 4 \(Mold Mitochondrial; Protozoan Mitochondrial; Coelenterate Mitochondrial; Mycoplasma; Spiroplasma\)](#)

[Lineage \(full\)](#)

cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Ctenophora; Tentaculata; Cydippida; Bathytectenidae; Bathytectena

External Information Resources (NCBI LinkOut)

| LinkOut                                          | Subject               | LinkOut Provider                                       |
|--------------------------------------------------|-----------------------|--------------------------------------------------------|
| <a href="#">Bathytectena chuni (Moser, 1909)</a> | taxonomy/phylogenetic | <a href="#">Ocean Biogeographic Information System</a> |
| <a href="#">Bathytectena chuni (Moser, 1909)</a> | taxonomy/phylogenetic | <a href="#">World Register of Marine Species</a>       |

**Entrez records**

| Database name            | Direct links |
|--------------------------|--------------|
| Nucleotide               | 56           |
| Protein                  | 55           |
| PubMed Central           | 4            |
| SRA Experiments          | 1            |
| Identical Protein Groups | 50           |
| BioProject               | 2            |
| BioSample                | 5            |
| Taxonomy                 | 1            |

# Cas d'étude : The Open Tree of Life

NCBI Taxonomy Browser

Search for:  as: complete name ☐ lock

Display: 3 levels using filter: none

**Bathytectena chuni**

Taxonomy ID: 1403704 (for references in articles please use NCBI:txid1403704)

current name

***Bathytectena chuni* (Moser, 1909)**

basonym: *Mertensia chuni* Moser, 1909

NCBI BLAST name: comb jellies

Rank: species

Genetic code: [Translation table 1 \(Standard\)](#)

Mitochondrial genetic code: [Translation table 4 \(Mold Mitochondrial; Protozoan Mitochondrial\)](#)

[Lineage \(full\)](#)

cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Ctenophora

External Information Resources (NCBI LinkOut)

| LinkOut                                          | Subject               |
|--------------------------------------------------|-----------------------|
| <a href="#">Bathytectena chuni (Moser, 1909)</a> | taxonomy/phylogenetic |
| <a href="#">Bathytectena chuni (Moser, 1909)</a> | taxonomy/phylogenetic |

WoRMS World Register of Marine Species

marine only ☐ extant only ☐

Home About Subregisters Users Photogallery Documents LifeWatch Contribute

Quick search:  Taxa Literature Distribution Specimen Editors Statistics Tools Manual Log in

### WoRMS taxon details

★ ***Bathytectena chuni* (Moser, 1909)**

AphiaID: 265844 (urn:lsid:marinespecies.org:taxname:265844)

Classification: Biota → ★ Animalia (Kingdom) → ★ Ctenophora (Phylum) → ★ Tentaculata (Class) → ★ Cydippida (Order) → ★ Bathytectenidae (Family) → ★ *Bathytectena* (Genus) → ★ *Bathytectena chuni* (Species)

Status: accepted

Rank: Species

Parent: ★ *Bathytectena* Mortensen, 1912

Original name: ★ *Mertensia chuni* Moser, 1909

Synonymised names: ★ *Mertensia chuni* Moser, 1909 · unaccepted > superseded combination

Environment: marine, brackish, fresh, terrestrial

Original description: (of ★ *Mertensia chuni* Moser, 1909) Moser, F. (1909). Die Ctenophoren der Deutschen Südpolar-Expedition 1901-1903. Deutsche Südpolar-Expedition 1901-1903 im Auftrage des Reichsamtes des Innern. E. von Drygalski (ed.). Berlin, Druck und Verlag von Georg Reimer. 1910: 117-192, pls XX-XXII. page(s): 127-130, pl. 20, figs. 1-4 (details)

Type data: ★ Note Südwestlich von Kerguelen (58° 26' S, Br. 90°...)

Taxonomic citation: Mills, C.E. Internet (1998-present). Phylum Ctenophora: list of all valid species names. Electronic internet document. *Bathytectena chuni* (Moser, 1909). Accessed through: World Register of Marine Species at: <https://www.marinespecies.org/aphia.php?p=taxdetails&id=265844> on 2025-04-07

[Ocean Biogeographic Information System](#)

[World Register of Marine Species](#)

# Cas d'étude : The Open Tree of Life

NCBI Taxonomy Browser

Search for:  Enter PubMed Nucleotide Protein

Display: 3 levels using filter: none as complete name lock

**Bathytectena chuni**

Taxonomy ID: 1403704 (for references in articles please use NCBI:txid1403704)

current name

***Bathytectena chuni* (Moser, 1909)**  
basonym: *Mertensia chuni* Moser, 1909

NCBI BLAST name: comb jellies

Rank: species

Genetic code: Translation table 1 (Standard)

Mitochondrial genetic code: Translation table 4 (Mold Mitochondrial; Protozoan Mitochondrial)

Lineage (full)

cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Ctenophora

External Information Resources (NCBI LinkOut)

| LinkOut                                          | Subject               |
|--------------------------------------------------|-----------------------|
| <a href="#">Bathytectena chuni (Moser, 1909)</a> | taxonomy/phylogenetic |
| <a href="#">Bathytectena chuni (Moser, 1909)</a> | taxonomy/phylogenetic |

WoRMS World Register of Marine Species

marine only extant only

Home About Subregisters Users Photogallery Documents LifeWatch Contribute

Quick search... Q Taxa Literature Distribution Specimen Editors Statistics Tools Manual Log in

## WoRMS taxon details

### ★ *Bathytectena chuni* (Moser, 1909)

|                      |                                                                                                                                       |
|----------------------|---------------------------------------------------------------------------------------------------------------------------------------|
| AphiaID              | 265844 (urn:lsid:marinespecies.org:aphis:265844)                                                                                      |
| Classification       | Biota → ★ <i>Animalia</i> (Kingdom) → ★ <i>Bathytectena</i> (Genus)                                                                   |
| Status               | accepted                                                                                                                              |
| Rank                 | Species                                                                                                                               |
| Parent               | ★ <i>Bathytectena</i> Mortensen, 1932                                                                                                 |
| Original name        | ★ <i>Mertensia chuni</i> Moser, 1909                                                                                                  |
| Synonymised names    | ★ <i>Mertensia chuni</i> Moser, 1909                                                                                                  |
| Environment          | marine, brackish, fresh, boreal                                                                                                       |
| Original description | (of ★ <i>Mertensia chuni</i> Moser, 1909) Expedition 1901-1903 in Australia 117-192, pls XX-XXII, page(s): 127-130, pl. 20, figs 1-10 |
| Type data            | ★ Note Südwestlich von Ker                                                                                                            |
| Taxonomic citation   | Mills, C.E. Internet (1998-present). Accessed through: WoRMS 2025-04-07                                                               |

Get data How-to Tools Community About

Classification

Select a species

Kingdom: Animalia

Phylum: Ctenophora

Class: Tentaculata

Order: Cyrtippida

Family: Bathytectenidae

Genus: Bathytectena Mortensen, 1932

Species: Bathytectena chuni (Moser, 1909)

No children

SPECIES | ACCEPTED

***Bathytectena chuni* (Moser, 1909)**

In: GBIF Backbone Taxonomy

OVERVIEW METRICS

10 OCCURRENCES

7 GEOREFERENCED RECORDS

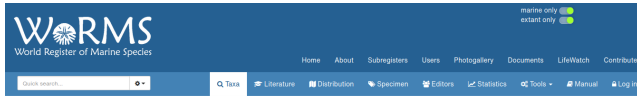
200 km

Generated by GBIF. OpenStreetMap contributors, © OpenMapTiles, GBIF

Year 2006

EXPLORE AREA

# Cas d'étude : The Open Tree of Life



**Bathytetena chuni**

Taxonomy ID: 1403704 (for references in articles)

current name

**Bathytetena chuni** (Moser, 1909)

basonym: **Mertensia chuni** Moser

NCBI BLAST name: **comb jellies**

Rank: **species**

Genetic code: [Translation table 1 \(Standard\)](#)

Mitochondrial genetic code: [Translation table 2 \(Vertebrate\)](#)

[Spiroplasma](#)

[Lineage \(full\)](#)

[cellular organisms](#); [Eukaryota](#); [Opisthokonta](#)

External Information Resources (NCBI LinkOut)

[LinkOut](#)

[Bathytetena chuni \(Moser, 1909\)](#)

[Bathytetena chuni \(Moser, 1909\)](#)

## WoRMS taxon details

### ★ *Bathytetena chuni* (Moser, 1909)

AphiaID 265944 (urn:lsid:marinespecies.org:aphia:265944)

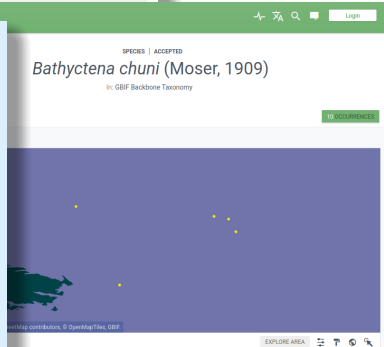


[Intro](#) | [Search taxa](#) | [Taxon tree](#) | [Search literature](#) | [Taxon match](#) | [Homonymy](#) | [Statistics](#) | [Webservice](#) | [Manual](#) | [FAQ](#) | [LifeWatch](#) | [Download](#) | [Log in](#)

## IRMNG taxon details

### *Bathytetena chuni* (Moser, 1909)

|                        |                                                                                                                                                                                  |         |            |
|------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------|------------|
| IRMNG_ID               | 11329585 (urn:lsid:irmng.org:taxname:11329585)                                                                                                                                   |         |            |
| Classification         | Biota -> Animalia (Kingdom) -> Ctenophora (Phylum) -> Tentaculata (Class) -> Cydippida (Order) -> Bathytetenoidea (Family) -> Bathytetena (Genus) -> Bathytetena chuni (Species) |         |            |
| Status                 | accepted                                                                                                                                                                         |         |            |
| Rank                   | Species                                                                                                                                                                          |         |            |
| Parent                 | Bathytetena Mortensen, 1932                                                                                                                                                      |         |            |
| Environment            | marine                                                                                                                                                                           |         |            |
| Fossil range           | recent only                                                                                                                                                                      |         |            |
| Original description   | Not documented                                                                                                                                                                   |         |            |
| Taxonomic citation     | IRMNG (2021). Bathytetena chuni (Moser, 1909). Accessed at: https://www.irmng.org/aphia.php?p=taxdetails&id=11329585 on 2025-04-07                                               |         |            |
| Taxonomic edit history | Date                                                                                                                                                                             | action  | by         |
|                        | 2006-09-20 22:00:00Z                                                                                                                                                             | created | Rees, Tony |
|                        | 2011-12-31 23:00:00Z                                                                                                                                                             | changed | Rees, Tony |
|                        | <a href="#">[taxonomic tree]</a>                                                                                                                                                 |         |            |



Deux caractéristiques sont ici à retenir : la *taille* de ce répertoire, que l'on devine gigantesque ; le degré de *standardisation*, tel que des publications qui incluraient une espèce donnée...donnent lieu à une même liste de références, incluant photos, ensemble des localisations, date de chaque photo, caractéristiques de l'espèce, mention dans un tableau comprenant toutes les occurrences de cette espèce. (Huneman 2023, p. 49)

# Les « voyages des données »

**data journeys:** the material, social, and institutional circumstances by which data are packaged and transported across research situations, so as to function as evidence for a variety of knowledge claims. (Leonelli 2016, p. 5)

**Mertensia Chuni** nov. sp.

Tafel XX fig. 1—4.

*Mertensia* CHUN p. 509 Textfigur, 1900.

Fundorte: 10. Februar: 1902: Südwestlich von Kerguelen ( $58^{\circ} 26'$  s. Br.  $90^{\circ} 22'$  ö. L.). Vertikal 2700 m; 1 Exemplar 30 mm.

MOSER, Ctenophoren.

127

9. Oktober 1903: Westlich von den Kap Verden ( $17^{\circ}$  n. Br.  $30^{\circ}$  w. L.). Vertikal 3000 m: 1 Exemplar 11 mm.

Das wertvollste Stück unter dem schönen Ctenophorenmaterial der Südpolar-Expedition war zweifelsohne ein großes, tadellos erhaltenes und konserviertes Exemplar einer Tiefsee-Ctenophore, welches der „Gauss“ das seltene Glück hatte, auf der Hinreise am 10. Februar 1902 unter dem  $58^{\circ} 26'$  s. Br. und dem  $90^{\circ} 22'$  ö. L., also südwestlich von Kerguelen, bei einem Vertikalzug von 2700 m zu erbeuten. Auf der Rückreise, viel weiter nördlich, schon auf der anderen Halbkugel verfieng sich ein zweites, jüngeres Exemplar am 9. Oktober 1903 im Vertikalnetz, westlich von den Kap Verden, unter dem  $30^{\circ}$  w. L. und dem  $10^{\circ}$  n. Br., ebenfalls bei einem Tiefzug, und zwar von 3000 m; leider war es ziemlich lädiert, stark geschrumpft und bespickt mit allem möglichen Getier; sehr wahrscheinlich war es schon tot, als es gefangen wurde, wofür das ganze Aussehen sprach.

# Les « voyages des données »

Yet for successful reuse to take place, the journey that data ultimately undertake should be determined as much by their users as it is by their curators. Curators cannot possibly predict all the ways in which data might be used. [...] Therefore, the best way to explore and maximize the value of data as evidence is to enable as many researchers as possible to use data in their own way and within their own research context. (Leonelli 2016, 26)

# L'analyse des données

---

Le *Big Data* ne diffère pas des simples données simplement en quantité ; c'est l'espace dans lequel ces données se déploient qui en vient à différer. Là où la donnée classique requiert un ou deux axes, elle peut en exiger des centaines dans le cas du *Big Data*, pour lequel l'espace des nuages de points où l'information pertinente doit être repérée devient un hyperespace à « n » dimensions, ce « n » pouvant indéfiniment croître. (Huneman 2023, p. 54)

# Changer la définition de la connaissance

Quelles sont leurs limitations ?

- Biaisé vers le présent (parfois mal archivé)
- Biaisé vers le texte (et contre image, vidéo, etc.)
- Biaisé vers le groupe (et contre l'individu)

Chaque méthode de recherche est limitée ! Il faut **mieux comprendre comment ces aspects changent notre recherche.**

# Plus de données $\neq$ meilleures données

Ajouter **plus de données** ne sert à rien si les données **ne sont pas une échantillon représentative de la population.**

## Exemple

Analyser plus de Tweets ne veut pas dire que l'image de « la politique belge » sur Twitter représente toute la population belge ! D'autant plus quand Twitter ne veut pas vous dire comment ils sélectionnent les tweets qu'ils fournissent aux chercheurs.

# Plus de données $\neq$ meilleures données

Si **plus de données** veut dire **intégrer des sources multiples**, ça entraîne d'autres problèmes.

## Exemple

Pour arriver à une bonne perspective d'une population biologique, comment intégrer données sur la distribution dans l'espace et le temps, la génétique, les observations « classiques » du comportement, etc...

# Données sans contexte

Les abstractions mathématiques ou statistiques qu'on analyse ne sont pas toujours équivalentes aux autres objets d'étude.

## Exemple

Si l'on tire le « réseau » d'une personne sur Facebook ou Twitter, c'est pas clair du tout comment c'est lié aux autres réseaux qu'on a étudié dans le passé : des relations familiales, ou au travail, ou le réseau social « en présentiel ».

# L'accès et l'éthique

Le fait de pouvoir utiliser des données ne veut pas dire que les personnes qui les ont créées consentent à leur utilisation pour tel ou tel projet de recherche.

Comment assurer que les projets respectent l'éthique ? Les scientifiques peuvent savoir des choses à propos de nous que nous ne savons même pas.

# Déclarations de l'objectivité incorrectes

Il reste **toujours des jugements subjectifs** dans l'interprétation des résultats, même si les données sont de qualité parfaite.

# Déclarations de l'objectivité incorrectes

- Qu'est-ce qui compte comme une donnée, qu'est-ce qui sera compté ou mesuré ?
- Quelles variables sont du « bruit » qui doit être « ignoré » ?
- Comment gérer les erreurs dans nos données, ou compenser pour des données manquantes ? Qu'est-ce qu'une « aberration statistique » ?

En écologie, ces dernières années un débat méthodologique a opposé certains chercheurs soutenant l'idée que la masse de données que nous récoltons, combinée au pouvoir nouveau du traitement statistique aidé par des algorithmes de type *machine learning*, permettrait de **prédire les changements** de la biodiversité ou de la composition des écosystèmes...sans construire une modélisation explicitement causale des processus en jeu – autrement dit, **sans comprendre** ce qui se passe dans le système. (Huneman 2023, p. 62)

**Il s'agit d'un monde où les quantités massives de données et les mathématiques appliquées remplacent tout autre outil susceptible d'être utilisé. Toutes les théories sur le comportement humain, de la linguistique à la sociologie, ont disparu. Autant pour toute théorie du comportement humain, de la linguistique à la sociologie. Oubliez la taxonomie, l'ontologie et la psychologie. Qui sait pourquoi les gens font ce qu'ils font ? [...] Avec suffisamment de données, les chiffres parlent d'eux-mêmes. (Anderson 2008)**

À quel type de connaissance correspond un savoir prédictif sans compréhension de « ce qui se passe » ? Sommes-nous revenus au fond à ce que Platon voyait dans l'astronomie, une manière de « sauver les phénomènes », autrement dit de décrire les orbites et prédire les positions des corps célestes sans aucune connaissance sous-jacente des processus et des mouvements ? Ou bien s'agit-il de quelque chose d'épistémologiquement nouveau ? (Huneman 2023, p. 67)

# Questions ?

charles@charlespence.net • <https://pencelab.be> •  @pence@scholar.social

