Domestiquer de nouvelles espèces de poissons grâce au text mining et à ISTEX

Mathieu Andro, Sophie Aubin

Domestiquer de nouvelles espèces de poissons grâce au text mining et à ISTEX

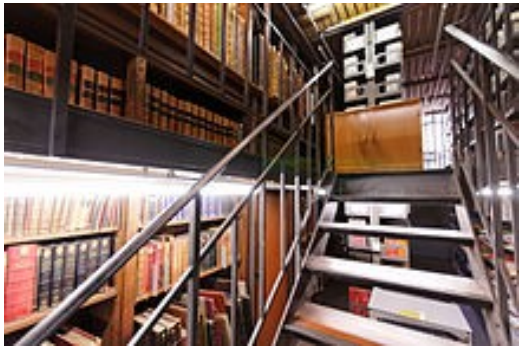
Mathieu Andro , Sophie Aubin (INRA, DIST)

Archives ouvertes et bases de publications : exploration et analyse des sources de données pour la recherche et ses environnements
IRHT, 23 mai 2016, 11 h 20 – 11 h 35

Objectifs

- *nvertir des textes en nnées*
- *crire le portrait robot du poisson qui peut être mestiqué*
- *Mieux comprendre le énomène de la domestication*
- *Rechercher les espèces qui lui ressemblent le plus (parmi 30 000 espèces connues)*

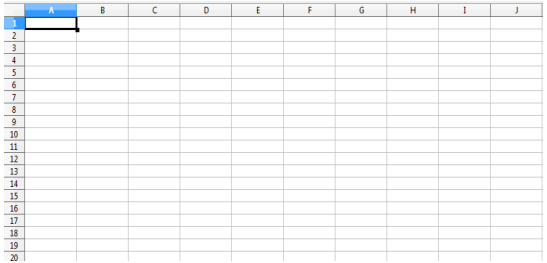
Hier



By Marie-Lan Nguyen - Own work, CC BY 2.5

CC BY-SA 3.0

Flickr, Kate Ter Haar, CC BY 2.0

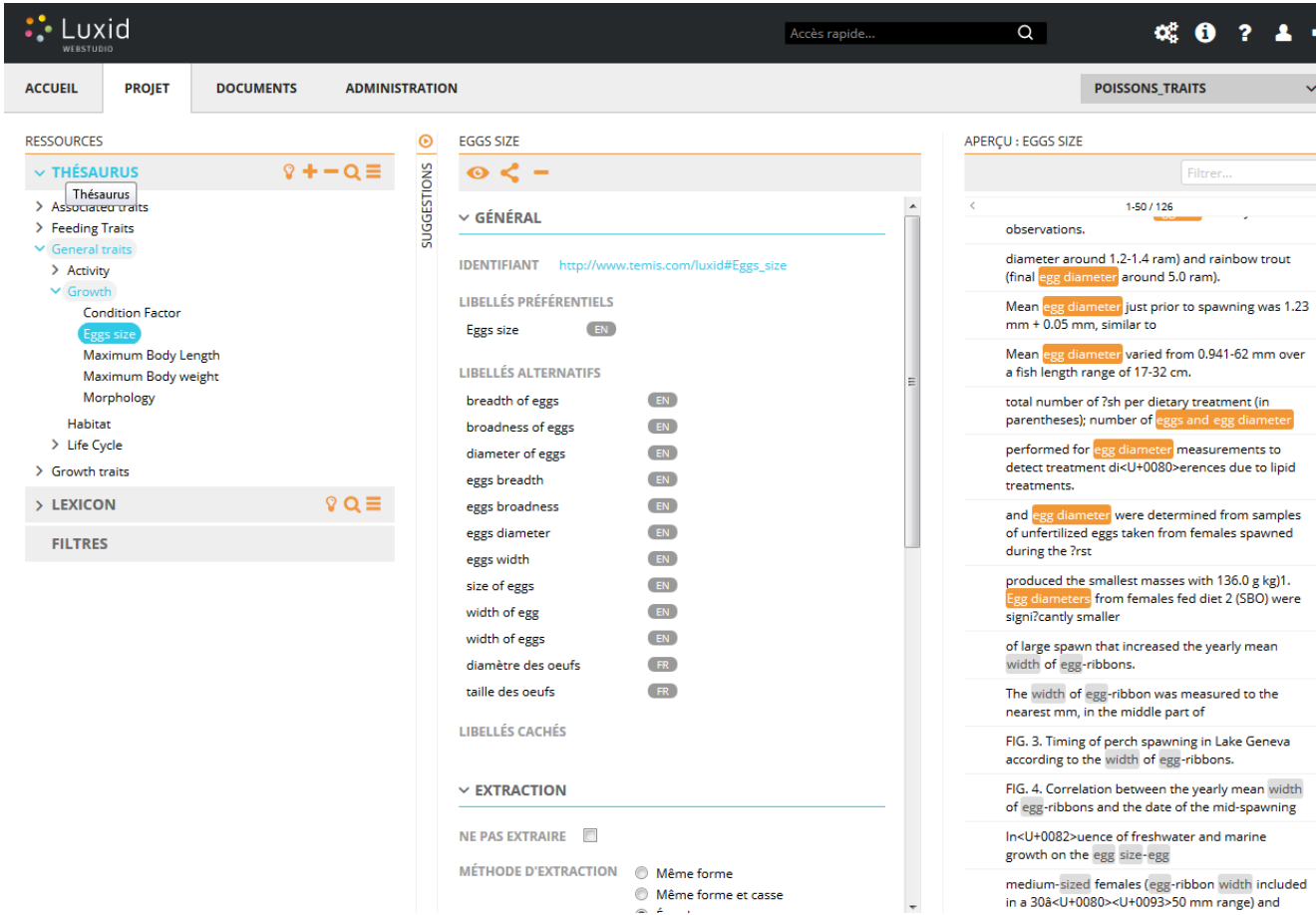


Aujourd'hui

- Constitution d'un corpus avec ISTEX
- Construction et utilisation de vocabulaires sur les traits de vie des poissons (reproduction, alimentation, croissance)
- Annotations grâce aux technologies de text mining
- Utilisation du crowdsourcing pour extraire et qualifier les données identifiées

- Modélisation des données

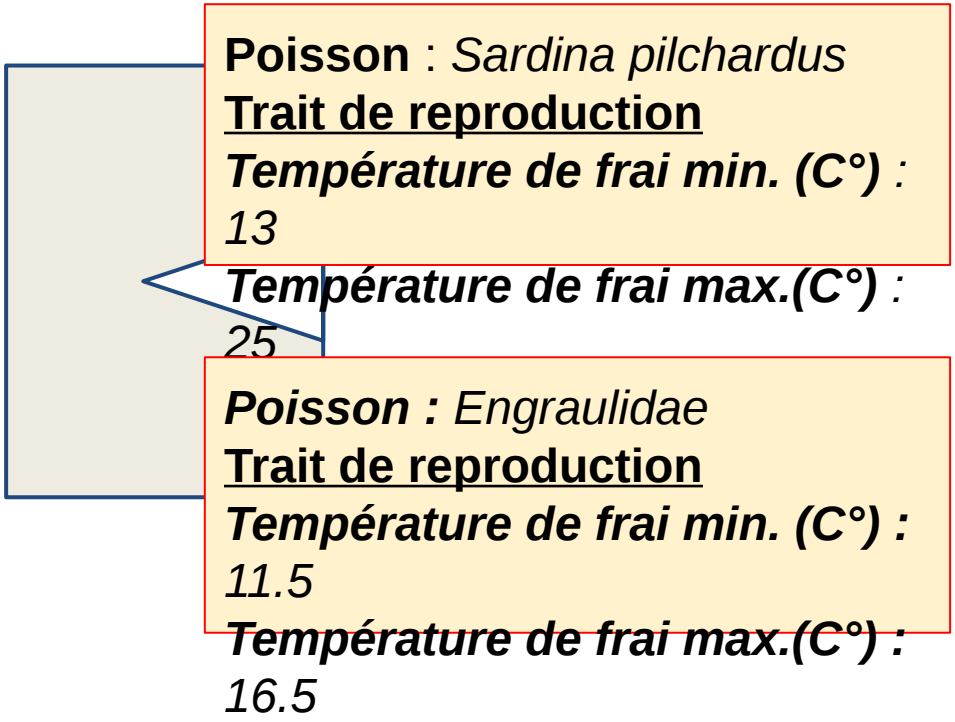
Construction de vocabulaires



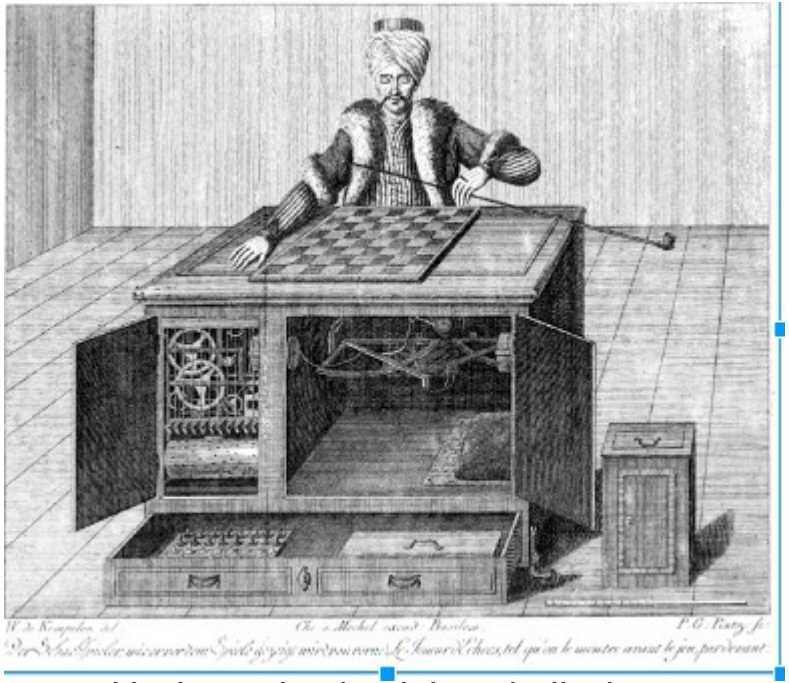
Transformer les textes en connaissances

"Sardines spawn in a much wider temperature range (13-25°C) than anchovy (11.5-16.5°C)."

(LLUCH-BELDA ET AL.: SARDINE AND ANCHOVY SPAWNING AS RELATED TO TEMPERATURE AND UPWELLING CalCOFI Rep., Vol. 32,1991)



Recours au crowdsourcing ?



« Tuerkischer schachspieler windisch4 » par Karl Gottlieb von Windisch. 1783. Public domain via Wikimedia Common

Extractions manuelles par crowdsourcing

Fecundity of perch estimated from a regression of egg number on fish length varied from 671 1 for a fish of 1 4 5 cm to 77978 for a female of 34.5 cm. Other independent variables related to egg number were age, gonad weight and somatic weight.

Mean egg diameter varied

from 0.94 to 1.62 mm. Males matured in their second year at a length of 6 cm and above while all females spawned at age four at a length of 15-1 8 cm. A proportion of females at Davan spawned one year earlier than at Kinord and were also slightly smaller on spawning. Faster growth of Davan fish may have encouraged earlier maturation. Spawning behaviour is described from film material.

- Zooplancton
- Nekton
- Others

What is the first feeding in hatchery (diet or live food) of the fish ?

Kind of food accepted by larvae at first feeding in hatchery

- artificial diet
- live food : artemia

What is the amount of food consummed daily at larval stages of the fish ?

Daily ration in larvae measured from larval weight in fed and starved state (Calculated % Body Weight)

What is the food or diet in juveniles and adult stages of the fish ?

Food items whose consumption has been observed in juveniles and adult stages

- Detritus
- Plants
- Zoobenthos
- Zooplancton
- Nekton
- Others

What is the amount of food consummed in juveniles and adults of the fish ?

Daily ration in juveniles and adults measured from amount of food voluntary consummed and whole body weight (calculated % Body Weight)

Morphology of the 1st fish mentionned in the text

What is the yolk sac length of the fish ?

Length of yolk sac along the main axis (mm)

ist@inra

Partenariats

UNIVERSITÉ DE LORRAINE

Wicri

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

FishBase

INRA SCIENCE & IMPACT

Recherches de financements

Demain

Flickr, Marc, Pescadería – Fish Shop, Madrid HDR, CC BY-NC-SA 2.0

Merci pour votre attention

— Contact : mathieu.andro@versailles.inra.fr

Diaporama : http://tinyurl.com/jytf55j

Bibliographie

- *Ben Ammar I, Teletchea F, Milla S, Ndiaye WN, Ledoré Y, Missaoui H, P. Fontaine (2015) Continuous lighting inhibits the onset of reproductive cycle in pikeperch males and females. Fish Physiology and Biochemistry 41: 345-356.*
- *FAO (2014) The State of World Fisheries and Aquaculture 2014 (SOFIA). 243 p. Rome : FAO.*
- *Fauconneau B. 2004 Diversification, domestication et qualité des produits. Productions AnimalesProd. Anim., 17 : 227-236.*
- *Fauconneau B. 2007 Flowthrough freshwater system in "Regional review on aquaculture developpment 6 Western-European region" Rana K.J. -2005. FAO Fisheries Circular N° 1017.6. Rome : FAO.*
- *Fauconneau, B., Teletchea, F., Andro, M., Mader, C., Le Bail, P.-Y., Bugeon, J., Geurden, I., Kaushik, S., Chatain, B., Baroillier, J.-F., Bardonnet, A., Begout-Anras, M.-L. (2014). Typologie des phénotypes nutritionnels chez les poissons d'élevage : constitution d'une base de méta-données. In: 4èmes Journées de la Recherche Filière Piscicole (p. 21). Presented at 4. Journées de la Recherche Filière Piscicole, Paris, FRA (2014-07-02 - 2014-07-04)*
- *Fostier A, Jalabert B (2004) Domestication et reproduction chez les poissons. Productions Animales, 17: 199-204.*
- *Pafilis E., et al. (2015). ENVIRONMENTS and EOL: identification of Environment Ontology terms in text and the annotation of the Encyclopedia of Life. Bioinformatics 31(11): 1872-1874.*
- *Pauly D, Christensen V, Dalsgaard J, Froese R, Torres FC (1998) Jr Fishing down marine food webs. Science. 1998, 279:860–863.*
- *Pauly D, Alder J, Bennett E, Christensen V, Tyedmers P, Watson R (2003) World The future for fisheries. Science. 2003, 302:1359–1361.*
- *Teletchea F, Fontaine P (2010). Comparison of early-life stage strategies in 65 European freshwater fish species: trade-offs are directed towards first-feeding of larvae in spring and early-summer. Journal of Fish Biology 77 : 257-278.*
- *Teletchea F, Fontaine P (2014) Levels of domestication in fish: implications for the sustainable future of aquaculture. Fish and Fisheries 15 : 181-195.*
- *Teletchea F, Fostier A, Kamler E, Gardeur JN, Le Bail PY, Jalabert B, Fontaine P (2009) Comparative analysis of reproductive traits in 65 freshwater fish species: application to the domestication of new fish species. Reviews in Fish Biology and Fisheries 19 : 403-430.*
- *Teletchea F, Fostier A, Le Bail PY, Jalabert B, Gardeur JN, Fontaine P (2007) STOREFISH: a new database dedicated to the reproduction of temperate freshwater teleost fishes. Cybium 31: 227-235.*

Résumé

Il y a une dizaine d'années, un scientifique avait extrait, "stabilo" en main, des données relatives à la reproduction des poissons des eaux tempérées au sein de la littérature imprimée et photocopiée. Dix ans plus tard, son travail de "fourmi" peut désormais être industrialisé et élargi avec l'aide des technologies de text mining et avec la mise à disposition du corpus ISTEX. Les données extraites de ce corpus permettront d'identifier les caractères types des poissons d'aquaculture afin de mieux comprendre le phénomène de la domestication mais aussi d'identifier les espèces de poissons qui ressemblent le plus à des espèces d'aquaculture, c'est à dire déjà domestiquées. Ces espèces identifiées pourront ainsi faire l'objet d'expérimentations de domestication afin de pouvoir, à l'avenir, diversifier les espèces de poissons d'aquaculture.

La présentation débutera par l'explicitation de la finalité sur le long terme du projet qui reste centrale : découvrir de nouvelles espèces de poissons à domestiquer au moyen des technologies de l'information scientifique et technique et plus particulièrement grâce au text mining et au corpus ISTEX. L'objectif plus immédiat du projet est ainsi de peupler une base de données relative aux caractères des espèces de poissons, en partenariat avec le consortium Fishbase, en ayant recours à des technologies de text mining. L'objectif de cette base de données sera ensuite de décrire le portrait robot de l'espèce domesticable et le portrait robot de l'espèce dont la domestication demeure impossible. Ce résultat permettra de mieux connaître le phénomène de la domestication des poissons (recherche fondamentale). Il permettra aussi d'identifier les espèces de poissons sauvages (30 000 espèces) qui ressemblent le plus à ces espèces de poissons d'aquaculture (une centaine d'espèces) afin de pouvoir en expérimenter la domestication (recherche appliquée).

La présentation du projet mettra en avant une comparaison entre la démarche suivie il y a plus de dix ans à la bibliothèque d'ichtyologie du Muséum national d'Histoire naturelle avec documents imprimés, photocopieur, stabylo et fichier Excel avec les outils de text mining actuellement utilisés afin de développer nos vocabulaires d'annotation. De la même manière, une comparaison entre nos négociations éditeur par éditeur sera effectuée avec l'API ISTEX actuellement utilisée. ISTEX nous a, en effet, permis de dépasser des obstacles aussi bien juridiques que techniques.

Ces comparaisons permettront, en outre, de mieux redonner du sens et de caractériser l'utilité concrète des technologies text mining. Elles aboutiront ensuite à un focus sur les technologies actuellement en cours de mobilisation autour des outils de text mining afin d'annoter des corpus de documents et de développer des vocabulaires à partir de corpus de textes.

Au delà de l'annotation d'un corpus ISTEX avec ces vocabulaires, le recours à du crowdsourcing est également envisagé afin de convertir des données annotées par les outils en données interprétées et validées par le cerveau humain et ainsi, "parcourir les derniers kilomètres" qui nous sépareront encore de la connaissance. L'analyse des données collectées et leur mise à disposition sous la forme d'un wiki sémantique, qui reste en grande partie encore à produire, sera également évoquées.

Enfin, les perspectives de partenariats avec le Muséum national d'Histoire naturelle et le consortium Fishbase seront abordées ainsi que les opportunités de financements qui ont été et restent encore à rechercher (appels à projets ANR, Interprofession, ISTEX, France AgriMer).