Petit retour d'expérience diachronique

Journée thématique transversale de l'ATILF
Ouvrir les données de la recherche : ressources, outils et retours d'expérience
Mardi 24 juin 2025

Alain Polguère – Université de Lorraine, CRNS, ATILF

ANALYSE ET TRAITEMENT INFORMATIQUE DE LA LANGUE FRANÇAISE

Objet limité de l’intervention

- Uniquement expérience personnelle
- Uniquement sur la distribution de ressources comme produits de recherche – sujet des publications scientifiques (presque) pas abordé
- Examen de deux expériences de distribution de données
  - 1er cas de figure : distribution des données de la base DiCo
    Début années 2000 : OLST, Université de Montréal
  - 2e cas de figure : distribution des données du *Réseau Lexical du Français*
    2013– : ATILF CNRS

Il faut qu’une science soit ouverte ou fermée

Au-delà du mot à la mode

- Open science is a global movement that aims to make scientific research and its outcomes freely accessible to everyone. By fostering practices like data sharing and preregistration, open science not only accelerates scientific progress but also strengthens trust in research findings.
  Adopting open science practices can enhance the quality, credibility, and reach of your research.
  Open science is a collaborative effort that welcomes everyone—regardless of role or experience—to participate in creating a more equitable and trustworthy research ecosystem.
  Center for Open Science – https://www.cos.io/open-science

- La science ouverte (ou *Open Science*) est un mouvement dont l’objectif est de rendre universellement accessibles les résultats de la recherche scientifique (publications et données de recherche, notamment). Concrètement, il s’agit de sortir ces connaissances des revues et des bases de données payantes ou fermées, pour les diffuser à tous – chercheurs, entreprises et citoyens – sans entrave, sans délai et gratuitement. L’Inserm soutient ce mouvement depuis 2003.
  INSERM – https://www.inserm.fr/nos-recherches/science-ouverte/

- La science ouverte consiste à rendre « accessible [sic] autant que possible et fermé [sic] autant que nécessaire » les résultats de la recherche, issus en majorité des fonds publics. Le CNRS est très engagé dans le développement de la science ouverte depuis de nombreuses années.
  CNRS – https://www.science-ouverte.cnrs.fr/science-ouverte/

1er cas : distribution de la base DiCo

Début années 2000 : OLST, Université de Montréal

- Accès **DiCo** via une interface de requête sur une base SQL : **Dicouèbe**
- Début des années 2000 – *Creative Commons* en est à ses débuts
- Principes adoptés en concertation avec l'administration de l'Université de Montréal
  - Liberté totale de consultation, extraction et exploitation
  - **MAIS** : s'assurer que les données ne seront pas victimes de **prédation** par un tiers, qui en verrouillera l'exploitation
  - Faire très simple
  - Philosophie du développement des logiciels libres (Amérique du Nord, années 80) + licence du WordNet de Princeton sont les exemples à suivre

Licence de WordNet 3.0 – Princeton University

WordNet License

This license is available as the file LICENSE in any downloaded version of WordNet.

WordNet 3.0 license: (Download)

WordNet Release 3.0 This software and database is being provided to you, the LICENSEE, by Princeton University under the following license. By obtaining, using and/or copying this software and database, you agree that you have read, understood, and will comply with these terms and conditions.: Permission to use, copy, modify and distribute this software and database and its documentation for any purpose and without fee or royalty is hereby granted, provided that you agree to comply with the following copyright notice and statements, including the disclaimer, and that the same appear on ALL copies of the software, database and documentation, including modifications that you make for internal use or for distribution. WordNet 3.0 Copyright 2006 by Princeton University. All rights reserved. THIS SOFTWARE AND DATABASE IS PROVIDED "AS IS" AND PRINCETON UNIVERSITY MAKES NO REPRESENTATIONS OR WARRANTIES, EXPRESS OR IMPLIED. BY WAY OF EXAMPLE, BUT NOT LIMITATION, PRINCETON UNIVERSITY MAKES NO REPRESENTATIONS OR WARRANTIES OF MERCHANT- ABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE OR THAT THE USE OF THE LICENSED SOFTWARE, DATABASE OR DOCUMENTATION WILL NOT INFRINGE ANY THIRD PARTY PATENTS, COPYRIGHTS, TRADEMARKS OR OTHER RIGHTS. The name of Princeton University or Princeton may not be used in advertising or publicity pertaining to distribution of the software and/or database. Title to copyright in this software, database and any associated documentation shall at all times remain with Princeton University and LICENSEE agrees to preserve same.

Licence du Dicouèbe (2005) – OLST, Université de Montréal

```
--------------------------------------------------------------------------------
FRANÇAIS (English follows)
--------------------------------------------------------------------------------

Base de données DiCo-OLST des dérivations sémantiques et collocations du français
Copyright (c) 2005 Observatoire de linguistique Sens-Texte (OLST). Tous droits réservés.

Les données contenues dans le DiCo-OLST sont octroyées aux LICENCIÉS
par l'Observatoire de linguistique Sens-Texte (OLST) selon la présente
licence. En consultant, utilisant et reproduisant ces données, vous
approuvez les conditions d'utilisation telles que vous les avez lues
et comprises.

La licence vous permet d'utiliser, de copier, de modifier et de
distribuer ces données ainsi que les documents du DiCo-OLST à toutes
fins et sans frais ni redevance. Une telle permission sous-entend
votre volonté à respecter les droits d'auteur, y compris les
avertissements, et ce, sur TOUTES les copies de données ou de documents
qui seront effectuées, y compris celles qui seront modifiées pour une
utilisation interne ou pour une distribution quelconque.

LES DONNÉES VOUS SONT FOURNIES PAR L'OLST EN L'ÉTAT, SANS AUTRES
GARANTIES OU DÉCLARATIONS, EXPLICITES OU IMPLICITES. À TITRE D'EXEMPLE,
SANS ÊTRE EXCLUSIF, L'OLST N'EFFECTUE NI LA REPRÉSENTATION NI LA GARANTIE
DE LA QUALITÉ MARCHANDE OU D'ADÉQUATION À UN USAGE OU BIEN DE LA
NON-VIOLATION DES BREVETS, DES DROITS D'AUTEUR, DES MARQUES DE COMMERCE
OU DE TOUT AUTRE DROIT DE TIERS PAR LES DONNÉES OU LES DOCUMENTS LICENCIÉS.

Le nom de l'OLST ne peut pas être utilisé dans le cadre de publicités
ou de campagnes publicitaires menant à la distribution des données ou
de bases de données dérivées. Le titre des droits d'auteur des données,
des bases de données dérivées et des documents doit, en tout temps,
rester la propriété de l'OLST et le LICENCIÉ s'engage à le maintenir ainsi.

[Pour télécharger la licence : http://idefix.ling.umontreal.ca/LICENCE.txt]
```

Résumé des fondements idéologiques adoptés

- Produits de la recherche – **non sensible** – financée par les **fonds publics** ne sont pas des produits commerciaux, mais sont un bien public et doivent être en accès entièrement libre
- Ces produits doivent répondre à des **standards de qualité** (i) scientifiques et (ii) de présentation/ structuration des données
- **Aucune restriction** n'est imposée sur l'utilisation, notamment, n'importe qui peut exploiter ces produits de recherche, **y compris à des fins commerciales**
- Unique limitation : produits de recherche restent la **propriété de l'entité scientifique qui les a développés**
  ⇒ on ne peut pas se les approprier et en limiter l'accès même une fois intégrés à un produit commercial – c'est un pis-aller, face à la **prédation**
- Approche qui (i) encourage la diffusion des idées et connaissances, (ii) maximise la qualité de la recherche – **tout le monde peut voir dans le détail la qualité, ou la médiocrité de ce qui est fait**

2e cas : distribution du *Réseau Lexical du Français*

2013– : ATILF CNRS

- 2011–2014 : projet **RELIEF**, financement Région Lorraine + FEDER (UE), en **partenariat industriel**
- Précautions prises pour anticiper la distribution des données du ***Réseau Lexical du Français*** (convention RELIEF), **avec le soutien du laboratoire – Jean-Marie Pierrel**
- Problèmes rencontrés avec le CRNS d'alors : *science ouverte* n'était pas dans son vocabulaire
  - Approche institutionnelle par défaut : chaque diffusion des données à un partenaire/utilisateur particulier doit faire l'objet d'une convention avec le CNRS
- Étienne Petitjean, bien que sympathique à la cause : « T'es un ayatollah (de la distribution libre) »
- Diffusion actuelle : sur **ORTOLANG** (Ollinger & Polguère), selon des principes stricts
  - **Séparation des résultats et des données de travail**
  - **Format de données minimaliste** – pas de recours à des normes de balisage type LMF (ISO)
  - **Documentation détaillée et soignée** (structure, terminologie, langue, mise en page)
    ⇒ Lien avec la **publication libre**

Recommandations

- Appliquer des standards de qualité clairement établis sur ce que doit être une distribution libre de données de recherche
  - Il ne suffit pas d'invoquer le terme à la mode (*buzzword*) *science ouverte*
- Exemple : documentation détaillée et de qualité, type rapport technique
- Qualité de la langue et des formats de présentation
- Ne faire confiance – sur le moyen terme – à **aucune** institutionnalisation
  Ce qui devient populaire tend à devenir une **cible de prédation** – par ex. :
  - *Open Science*
  - *Creative Commons*