# Traitement Automatique des Langues, Biens Communs Informationnels et Industries de la Langue

Gaël de Chalendar



CEA LIST
Laboratoire Vision et Ingénierie des Contenus

22/11/2014

# Sommaire

Constat :

- Grande diversité de logiciels de Traitement Automatique des Langues (TAL) publiés sous licence libre

Questions courantes :

- Gratuits, alors moins bon ?
- Code source mis à disposition, alors pas d'activité commerciale ?

Au contraire

- Participe à l'amélioreration rapide de la qualité et la fiabilité des logiciels de TAL
- Divers modèles commerciaux (business models)

**Conviction**

Les progrès scientifiques et techniques se trouvent accélérés par l'ouverture du code source, pour l'ensemble de la communauté et par conséquent pour les acteurs individuels.

# Sommaire

# Définitions des Biens Communs

Les éléments de définition et de réflexion autour des Biens Communs Informationnels évoqués ici proviennent en grande partie du livre Cause Commune de Philippe Aigrain, 2005

# Définitions des Biens Communs

## "Le" Bien Commun, St Thomas d'Aquin, d'après A. Modde(1949)

Trois notions :

# Définitions des Biens Communs

## "Le" Bien Commun, St Thomas d'Aquin, d'après A. Modde(1949)

Trois notions :

1. Tous les êtres sont pénétrés par l'influence transcendante de Dieu. Ils sont tous unis les uns aux autres comme les parties d'un même tout, et ce tout est l'univers

# Définitions des Biens Communs

## "Le" Bien Commun, St Thomas d'Aquin, d'après A. Modde(1949)

Trois notions :

1. Tous les êtres sont pénétrés par l'influence transcendante de Dieu. Ils sont tous unis les uns aux autres comme les parties d'un même tout, et ce tout est l'univers
2. Bien commun intelligible et céleste donné à l'Homme au travers de l'Église

# Définitions des Biens Communs

## "Le" Bien Commun, St Thomas d'Aquin, d'après A. Modde(1949)

Trois notions :

1. Tous les êtres sont pénétrés par l'influence transcendante de Dieu. Ils sont tous unis les uns aux autres comme les parties d'un même tout, et ce tout est l'univers
2. Bien commun intelligible et céleste donné à l'Homme au travers de l'Église
3. Le Bien commun de la cité pour se conformer à la volonté divine

# Définitions des Biens Communs

## "Les" Biens Communs, Philippe Aigrain

"Toute « chose » ou entité immatérielle à laquelle on a décidé de donner un statut de propriété commune, de la faire appartenir à tous, parce qu'elle n'appartient à personne"

# Biens Communs Informationnels (BCI)

- Peuvent être créés, échangés et manipulés sous forme d'information
- Les outils de création et le traitement sont souvent eux-mêmes informationnels
- Coût marginal de la duplication des BCI quasiment nul
- Valeur non réduite par le fait qu'un autre s'en serve, mais plus souvent augmentée par la faculté d'échange ou de communication accrue
- Exemples : connaissances, créations dans tous les médias, idées, logiciels

  

# La Tragédie des Communs

## Garrett Hardin, 1968

- Étude des paturages mis en commun en angleterre
- Fragilité des biens communs physiques quand :
  - Usage croissant (démographie)
  - Quête du profit
  - Érosion d'un système de valeurs communes
- Nécessité d'une économie administrée ou de la propriété individuelle

# Réfutation

## Elinor Ostrom et Charlotte Hess,

### Digression à propos de la licence CC-BY

Un paragraphe de l'article est réécrit, volontairement sans guillemets, à partir d'un texte sous CC-BY, bien évidemment cité. Un relecteur a douté du bien fondé de cette manière de faire. Nous nous sommes permis de contacter l'auteur, Hervé Le Crosnier, pour avoir son avis. Il a conforté notre analyse.

# Réfutation

## Elinor Ostrom et Charlotte Hess,

- Modèle de Hardin $\neq$ Communs réels
- Pas des ressources mais des lieux de négociations
- Individus communicants certains motivés par intérêt commun
- Pour BCI : non soustractibles

  

# Logiciels : un parfait exemple de BCI

## Logiciels Libres

Diffusés selon une licence assurant le maintien des 4 libertés :

- Utiliser
- Copier
- Étudier et modifier
- Redistribuer

# Sommaire

# Nécessité des BCI en TAL

- Langage : autant un Bien Commun que l'eau, l'air, etc.
- Fondamental pour les libertés publiques : réseaux sociaux, ciblage commercial, surveillance par les autorités civiles et militaires
- Danger des brevets et du code privateur
- Pas tous les logiciels mais toutes les technologies

# Variété de l'offre

## Des systèmes de tous types

- Implémentations d'une seule technologie, d'une tâche
- Plateformes génériques (du découpage en tokens à l'analyse sémantique et pragmatique)
- Monolingues ou multilingues
- En tant que fonctionnalité interne d'une application

# Ressources linguistiques

## Une situation moins bonne

- Assez grand nombre de corpus librement téléchargeables ou facilement accessibles, mais non libres
- Apparition plus récente de corpus annotés vraiment libres
- Assez peu de langues et surtout en anglais

Quelques exemples de corpus annotés en morphosyntaxe :

- ANC : seulement des textes librement diffusables depuis 2006 $\Rightarrow$ OANC
- Sequoia : de qualité mais encore relativement petit
- Free French Treebank : beaucoup plus grand mais à améliorer

# Sommaire

# Justification des BCI d'un point de vue économique

- "Les communs sont avant tout des lieux de négociations (il n'y a pas de communs sans communauté), gérés par des individus qui communiquent, et parmi lesquels une partie au moins n'est pas guidée par un intérêt immédiat, mais par un sens collectif."
- Mais : société capitaliste où la concurrence est (censée être) "libre et non faussée"
- Démontrer l'efficacité économique des BCI
- Idéal devant être à l'origine de mécanismes efficaces pour la création de richesse au niveau de l'entreprise.

# GATE, Université de Sheffield

## International

- LGPL
- Customisation, formation, développements à façon
- Sponsoring, partenariats privilégiés
- Clients : UK National Health Service, UK Food and Environment Research Agency, Public Health England, nombreuses PMEs et statups
- Nombreux utilisateurs “anonymes” ou par l’intermédiaires d’autres fournisseurs de services
- D’autres gros clients sous NDA

# FreeLing, Université Polytechnique de Catalogne

## International

- Double licence commerciale et GPL, passage en cours à AGPL
- Utilisé par de nombreuses sociétés, souvent pour fournir des services en mode SaaS
  - http ://tecnolecto.com/sentilecto - Moteur d'analyse d'opinion
  - http ://store.apicultur.com/ - Pay-per-use web APIs basées sur freeling
- Double licence pour l'intégration dans des logiciels propriétaires
  - Ruby Reader - application iPhone aidant des locuteurs japonais à comprendre des textes en anglais
  - MediaTuners - assistant automobile contrôlé par la voix (musique, appels, agenda, navigation, etc.)
- Mise en place en cours en interne d'une offre SaaS de freeling et d'autres outils de TAL

# Divers outils de l'INRIA Alpage

## En France

- Distribués sous licences LGPL et Cecill
- Collaborations avec plusieurs industriels :
  - Vera : traitement multilingue de parties libres de formulaires
  - Kwaga : traitement de mails
  - Viavoo : traitement d'avis de consommateurs
  - Lingua et Machina : extraction de terminologie et acquisition de réseaux lexicaux dans des domaines spécialisés

# LIMA, CEA LIST

## En France

- AGPL depuis janvier 2014 + propriétaire
- Adaptation encore en cours pour utiliser des ressources linguistiques libres
- Disponible en français et en anglais
- Modules et ressources propriétaires pour allemand, arabe, chinois, espagnol...
- Utilisé par ANT'inno au cœur de l'outil de veille ANT'box. Clients :
  - Eurocopter, Dipta, Inéris, Sodiaal, Cyrus
  - IRSN, Gouvernement Mauritanien, Ministères au Maroc et au Cameroun
- Support et services assurés par ANT'inno en collaboration avec le LVIC

# Autres exemples

- Stanford PoS tagger, Parser, NER : commercial licence available
- LingPipe (Thomson, Endeca, NLM, DARPA, MITRE)
- RapidMiner était libre mais est redevenu propriétaire (et forké)
- Gensim supporté par sa communauté. support commercial et développements spécifiques par son auteur principal
- Carrot2 (Open Source Search Results Clustering Engine) intégré par Carrot Search dans son produit commercial Lingo3G

  

# Sommaire

- Le langage est un des Biens Communs Informationnels les plus importants à préserver
- Fondamental de disposer de logiciels de TAL et de ressources linguistiques libres pour toutes les langues
- Des industriels s'engagent dans leur développement, leur exploitation et leur diffusion
- Les modèles économiques le permettant sont nombreux

  