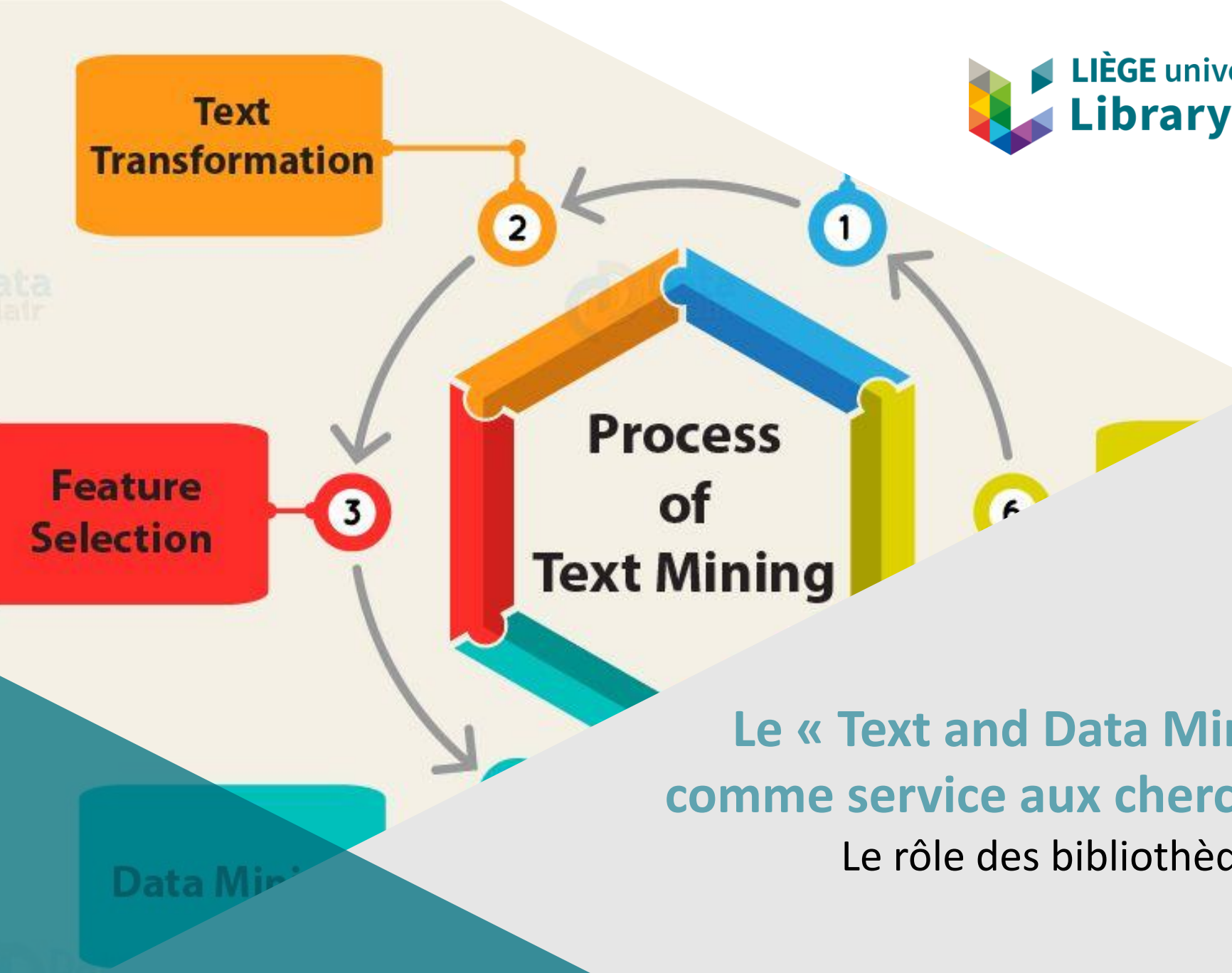
LIÈGE université
Library



Le « Text and Data Mining » comme service aux chercheurs

Le rôle des bibliothèques

0. Plan de l’exposé

1. Définition des termes
2. Aspects techniques
3. Aspects juridiques
4. Aspects scientifiques
5. Offre logicielle
6. Le rôle des bibliothèques
7. Conclusion

1. Définition des termes

Un flou artistique

1. Définition des termes

1.1 Définition des termes

- Un foisonnement lexical
  - Termes anglais, néologismes français, mots chapeaux, recoupements lexicaux, confusion entre des pratiques, des usages, des techniques précises, etc.
  - Text Mining, Data Mining, Web Mining, Citation Mining, etc.
  - Variation très grande selon la littérature consultée.
- Première résistance au concept : flou

1.1 Définition des termes

- Un foisonnement conceptuel
  - Approche historique du concept, approche par écoles et par domaines, approche par outils ou approche par usage, etc.
  - Croisement très grand de domaines : statistique, informatique, linguistique + domaines concernés.
- Deuxième résistance au concept : transdisciplinaire

1. Définition et historique des termes

- Un foisonnement technologique
  - plusieurs langages de programmation, plusieurs programmes, plusieurs API, plusieurs interfaces, etc.
  - Multiplication des technologies – et donc des usages.
- Troisième résistance au concept : vaste – et informatique
- Quatrième résistance : côté *buzzword*

1.2 Définition du *data mining*

- Qu'est-ce que le *data mining* ?
  - Le data mining est un processus d'extraction de structures inconnues, valides et potentiellement exploitables dans les bases de données, à travers la mise en œuvre des techniques statistiques et de machine learning (Fayyad 1996)
  - Ex. Econométrie, épidémiologie, *market analysis*, etc.

1.3 Définition du *text mining*

- Qu'est-ce que le *text mining* ?
  - L'application des techniques de *data mining* à un texte, après avoir rendu celui-ci exploitable par lesdites techniques.
  - Ex. lexicométrie, analyse des cooccurrences, analyse du style, analyse des opinions, résumé automatique, revue de la littérature automatique, identification automatique de concepts, etc.

1.3 Définition du *text mining*

Data Mining	Text Mining
Traitement immédiat	NLP (Natural Language Processing)
Identifie des relations causales	Découvre des informations inconnues
Données structurées	Données semi-structurées ou non structurées
Données numériques structurées souvent de même nature enregistrées dans des entrepôts de données	Données hétérogènes issues de corpus et de formats très éclectiques.

(Basé sur Ray 2017)

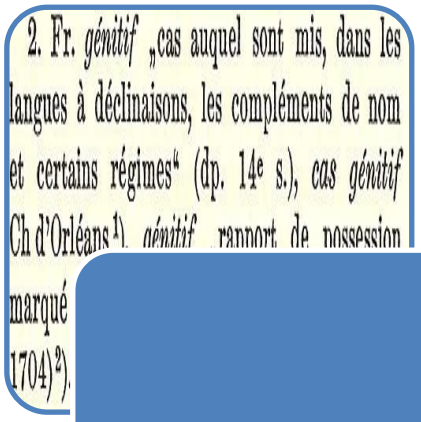
The journey from unstructured text to structured content

Text mining more specifically has been defined as the process of "structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and final evaluation and interpretation of the output (Libguide, University of Cambridge)

2. Aspects techniques

Étapes de travail et processus

2.1 Processus de travail

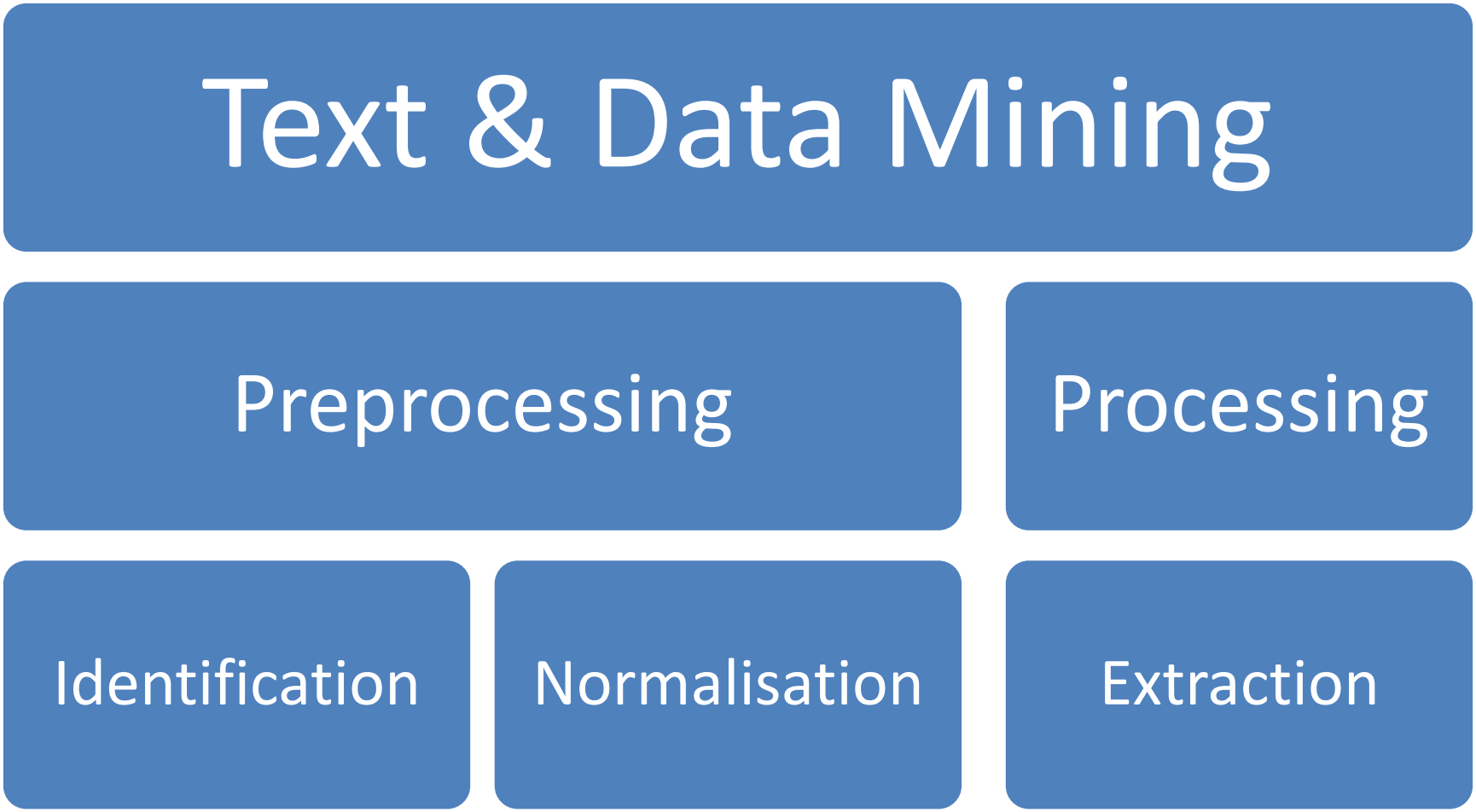


Texte brut

Données semi structurées

Données structurées

2.1 Processus de travail



2.2 Identification

Corpus is a (large) collection of text documents that has been brought together according to a certain set of predetermined criteria, in machine readable form and can be used for extracting statistical and linguistic information

2.3 Normaliser les données

- Pour exploiter le corpus, il faut normaliser (ou standardiser) les documents.
  - Océrisation du texte
  - Résolution des abréviations, symboles, etc.
  - Multiples opérations linguistiques
- La normalisation des données est parfois nécessaire pour le Data Mining.
  - Ex. Normalisation de données en *Unimarc*

2.3 Normaliser les données

- Plusieurs opérations sont nécessaires :
  - Nettoyage (cleaning & parsing)
  - Tokenisation
  - Filtrage divers
  - Lemmatisation (stemming/lemmatisation)

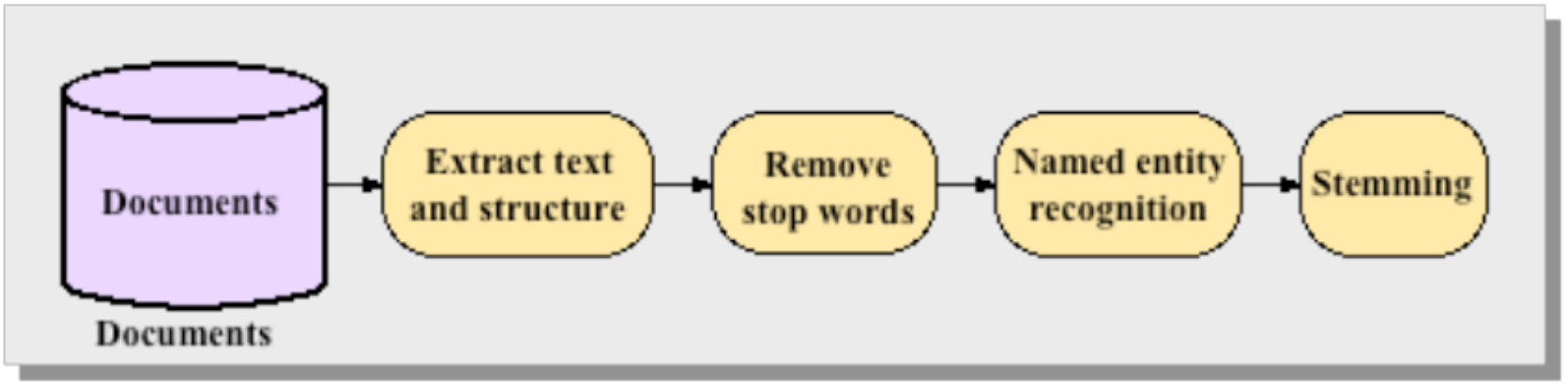
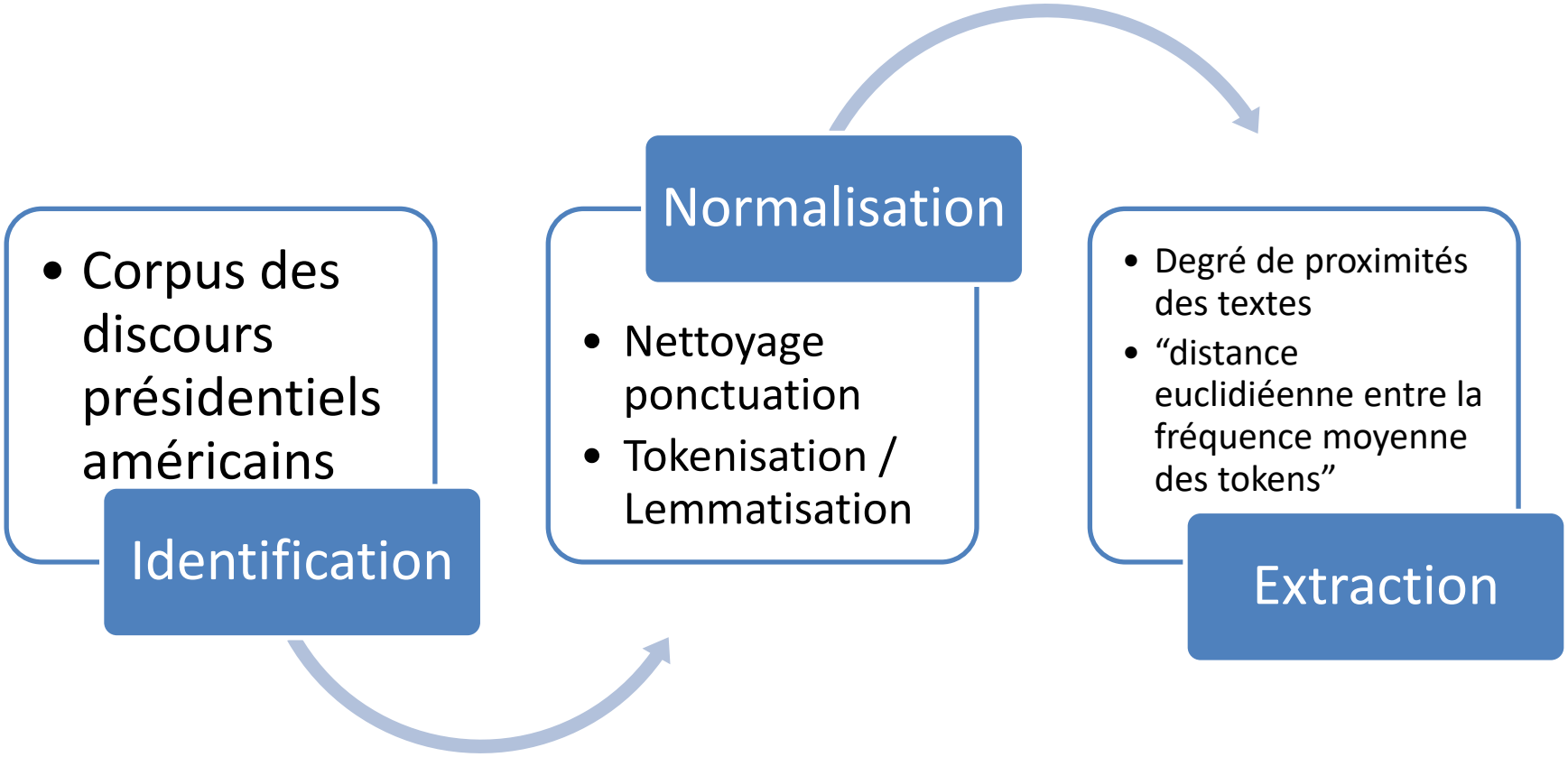


Fig. 1. Preprocessing (Mantrach et al.)

2.4 Extraction de données

- Confusions nombreuses autour du terme « *extraction* ».
  - « processus **d'extraction** de structures inconnues, valides et potentiellement exploitables »
- Création de nouvelles données sur la base des données étudiées.

2.5 Résumé


Identification
• Corpus des discours présidentiels américains
Normalisation
• Nettoyage ponctuation
• Tokenisation / Lemmatisation
Extraction
• Degré de proximités des textes
• “distance euclidiéenne entre la fréquence moyenne des tokens”

2.5 Résumé

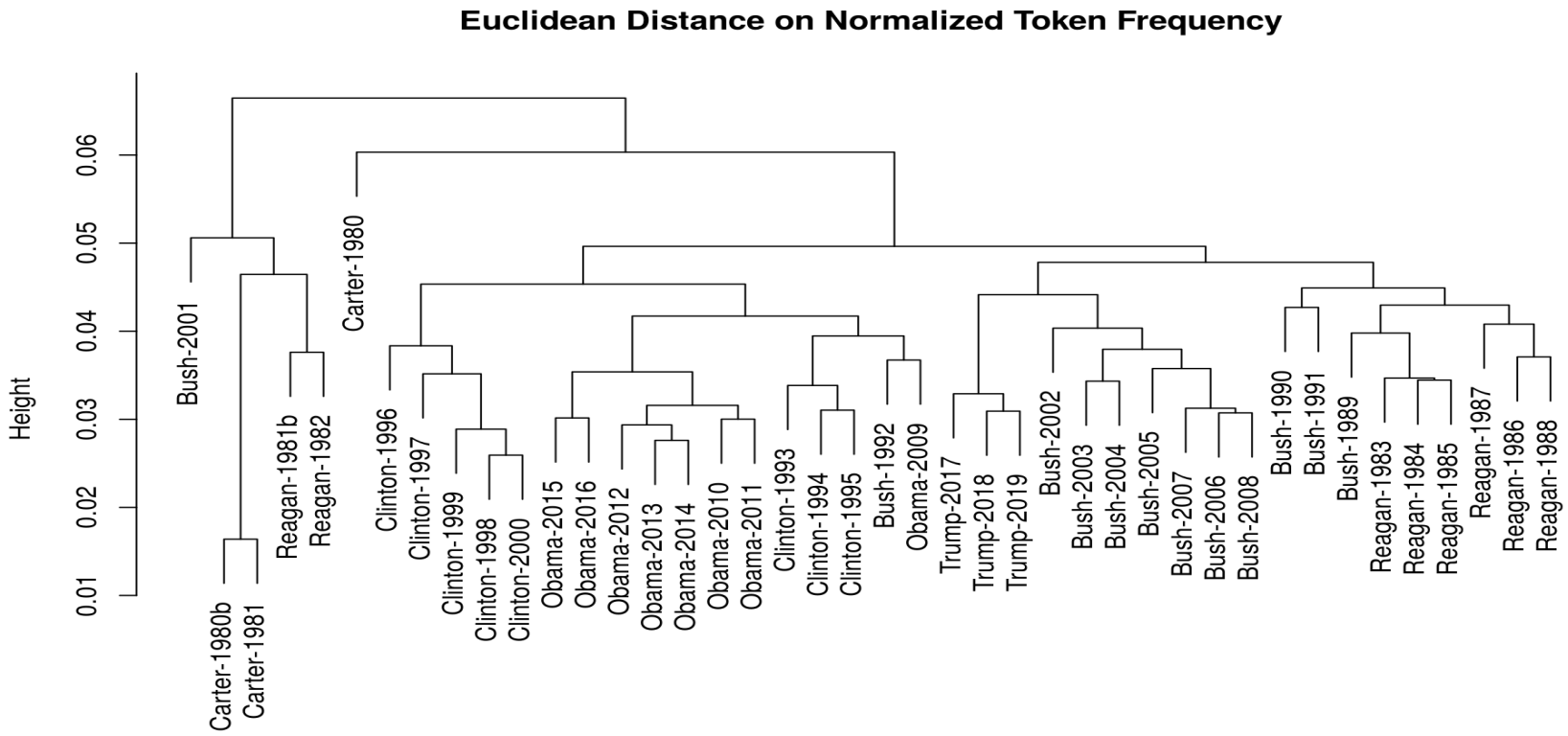


Fig. 2 Quanteda Quick Start Guide

2.6 Structure d'un processus

- Programme
  - Effectue la tâche demandée
- Les données à l'entrée (*input*)
  - Les données (structurées ou non) que nous souhaitons traiter.
- Les résultats en sortie (*output*)
  - Les données une fois traitées.
- Les ressources
  - Données auxiliaires qui permettent la réalisation de la tâche.

2.6 Structure d'un processus

- Programme
  - Nettoyage de la ponctuation.
- Les données à l'entrée (*input*)
  - Le corpus des textes présidentiels.
- Les résultats en sortie (*output*)
  - Le corpus sans la ponctuation.
- Les ressources
  - La liste de tous les symboles de ponctuation.

2.6 Structure d'un processus

- Programme
  - Nettoyage des *stop words*
- Les données à l'entrée (*input*)
  - Le corpus des textes présidentiels sans ponctuation.
- Les résultats en sortie (*output*)
  - Le corpus sans les *stopwords*.
- Les ressources
  - La liste de tous les *stopwords.*

2.6 Structure d'un processus

- Programme
  - Tokenisation
- Les données à l'entrée (*input*)
  - Le corpus des textes présidentiels sans ponctuation et sans stopwords.
- Les résultats en sortie (*output*)
  - Le corpus tokenisé
- Les ressources
  - La liste des « caractères » qui séparent les tokens.

2.6 Structure d'un processus

- Programme
  - Lemmatisation
- Les données à l'entrée (*input*)
  - Le corpus des mots triés par textes
- Les résultats en sortie (*output*)
  - Le corpus des lemmes triés par textes
- Les ressources
  - Dictionnaire lemmatiseur

2.6 Structure d'un processus

- Programme
  - Création d'un dataset
- Les données à l'entrée (*input*)
  - Corpus de listes des lemmes (répétés)
- Les résultats en sortie (*output*)
  - Un document avec les fréquences, etc.
- Les ressources
- /

2.6 Structure d’un processus

- Programme
  - Création d’un wordcloud
- Les données à l’entrée (*input*)
  - Dataset
- Les résultats en sortie (*output*)
  - Wordcloud
- Les ressources
  - /

2.6 Structure d’un processus

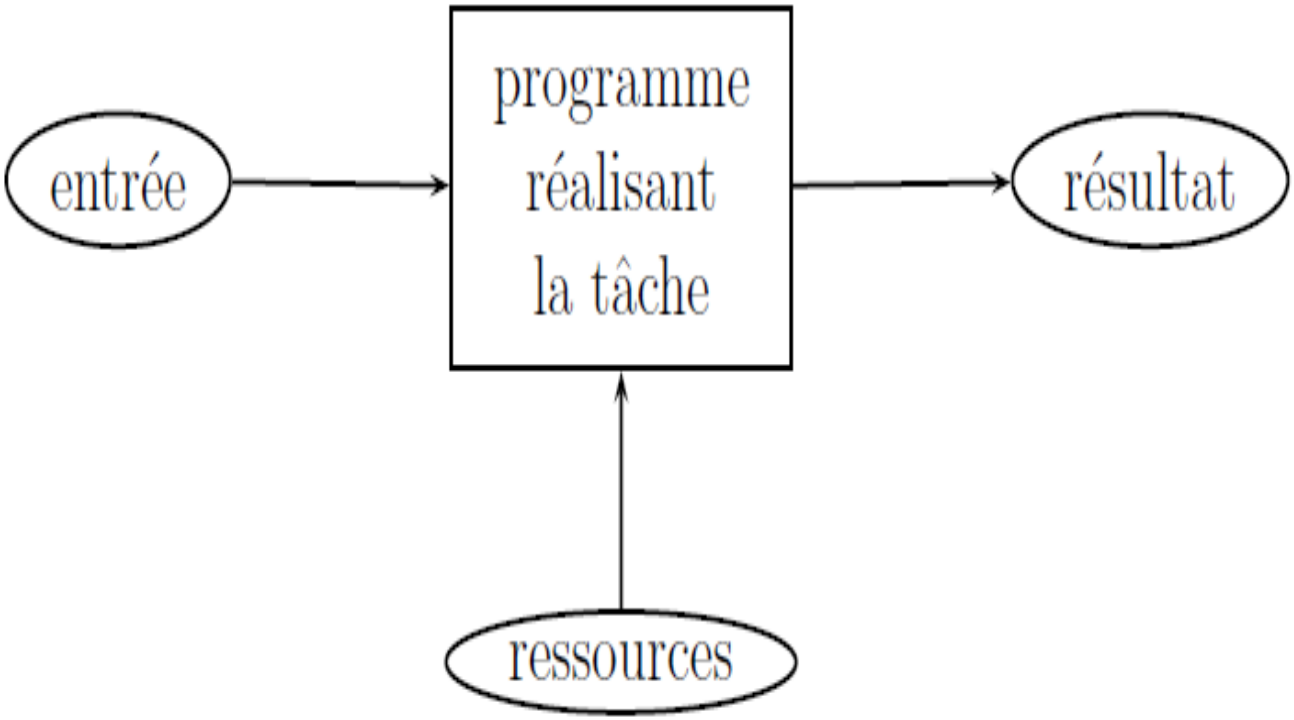
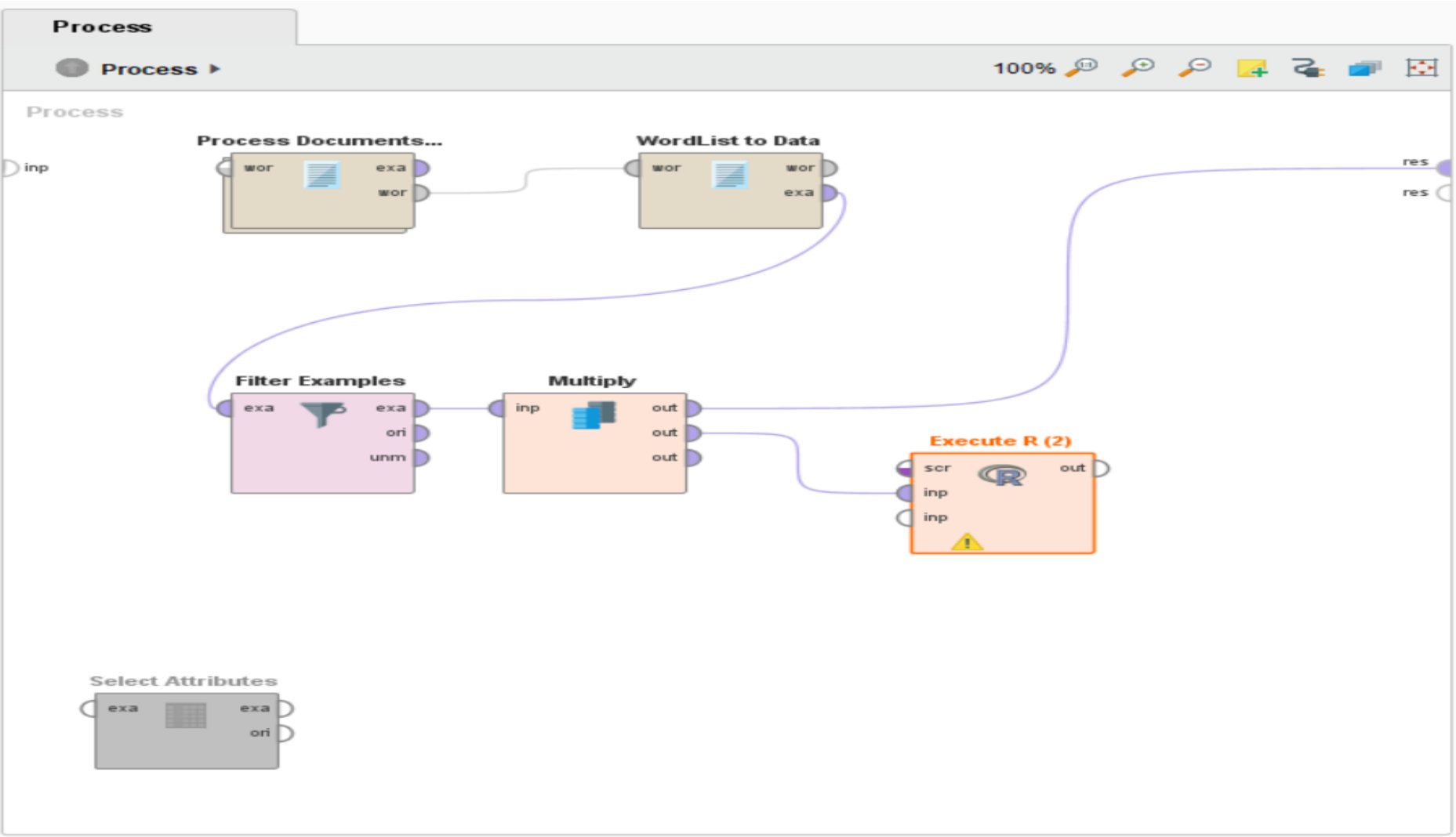


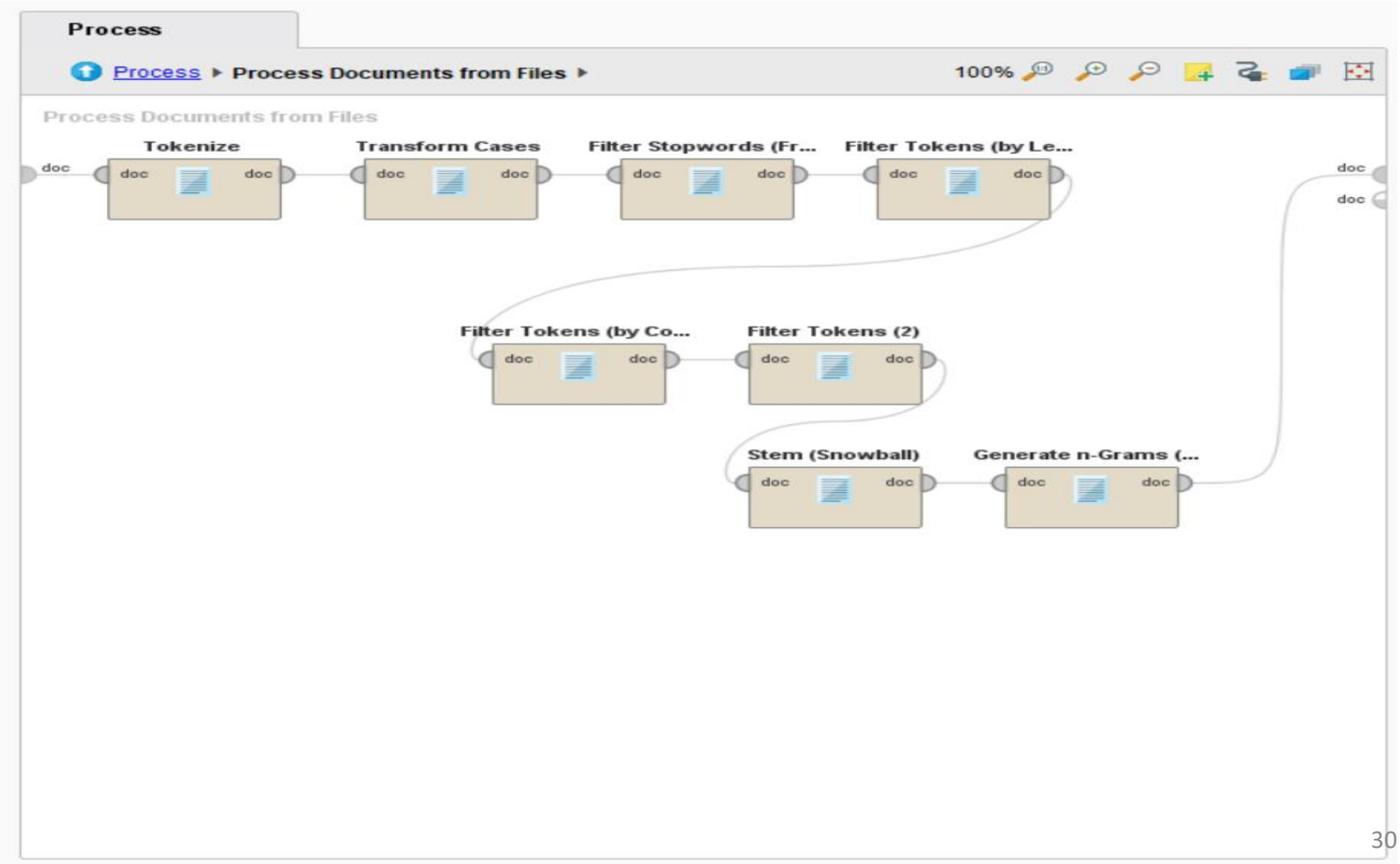
Figure 2 : Description d'une tâche (Tellier 2016)

Ce type de figure est un standard utilisé par de nombreux logiciels et ressources (OpenMinTed, Rapidminer, et.)

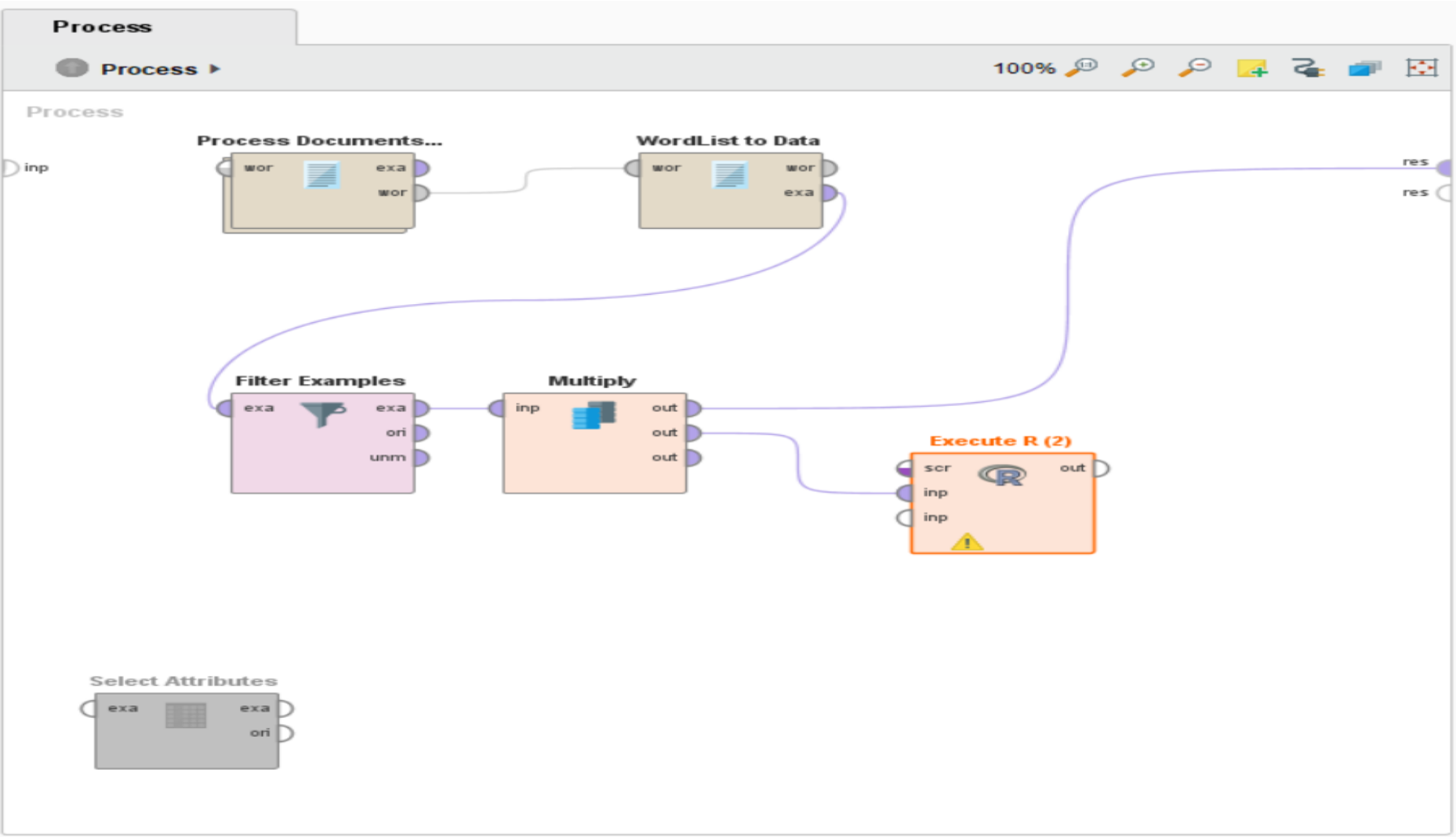
2.6 Structure d’un processus



2.6 Structure d’un processus



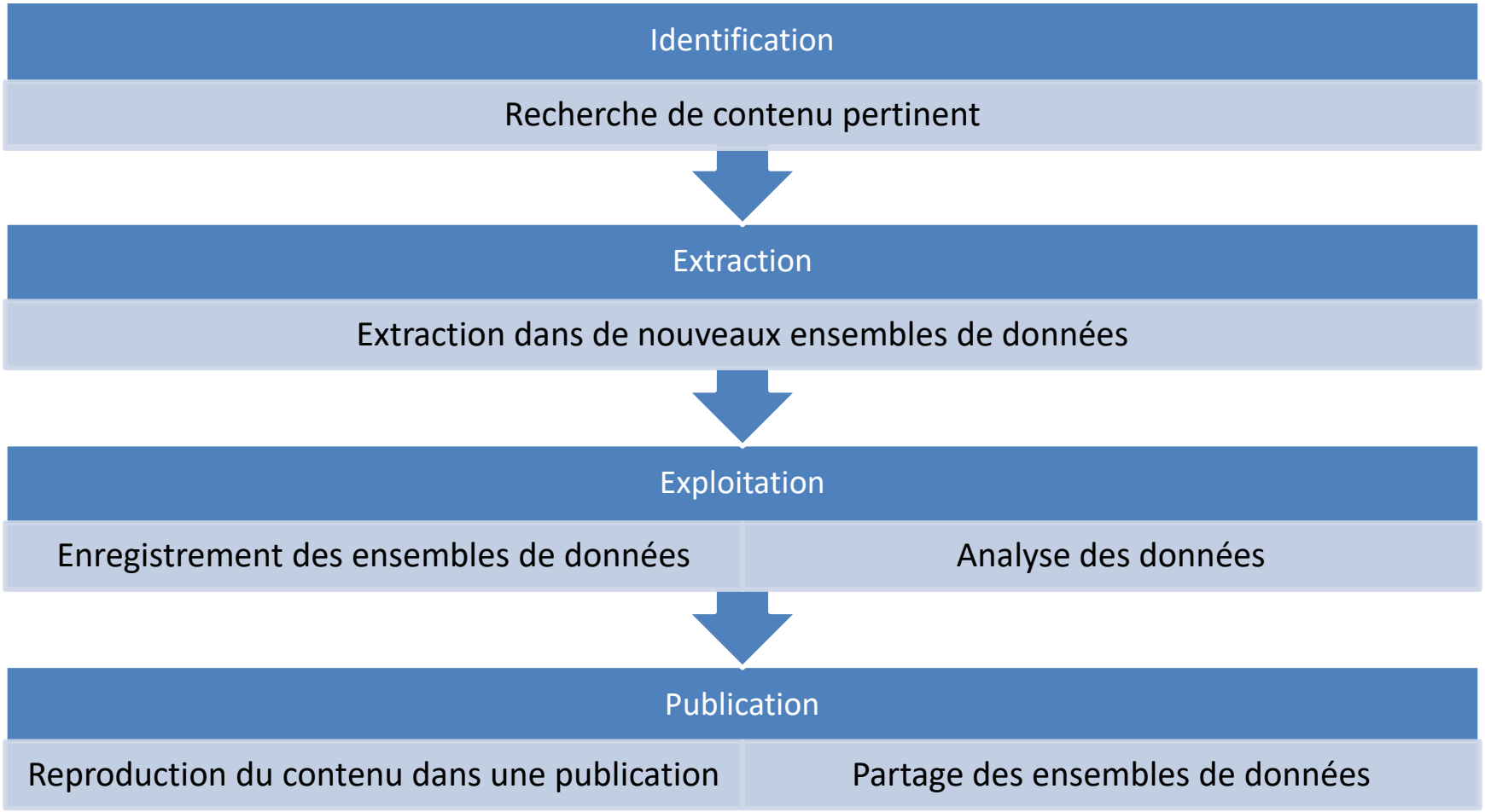
2.6 Structure d’un processus



LIÈGE université Library

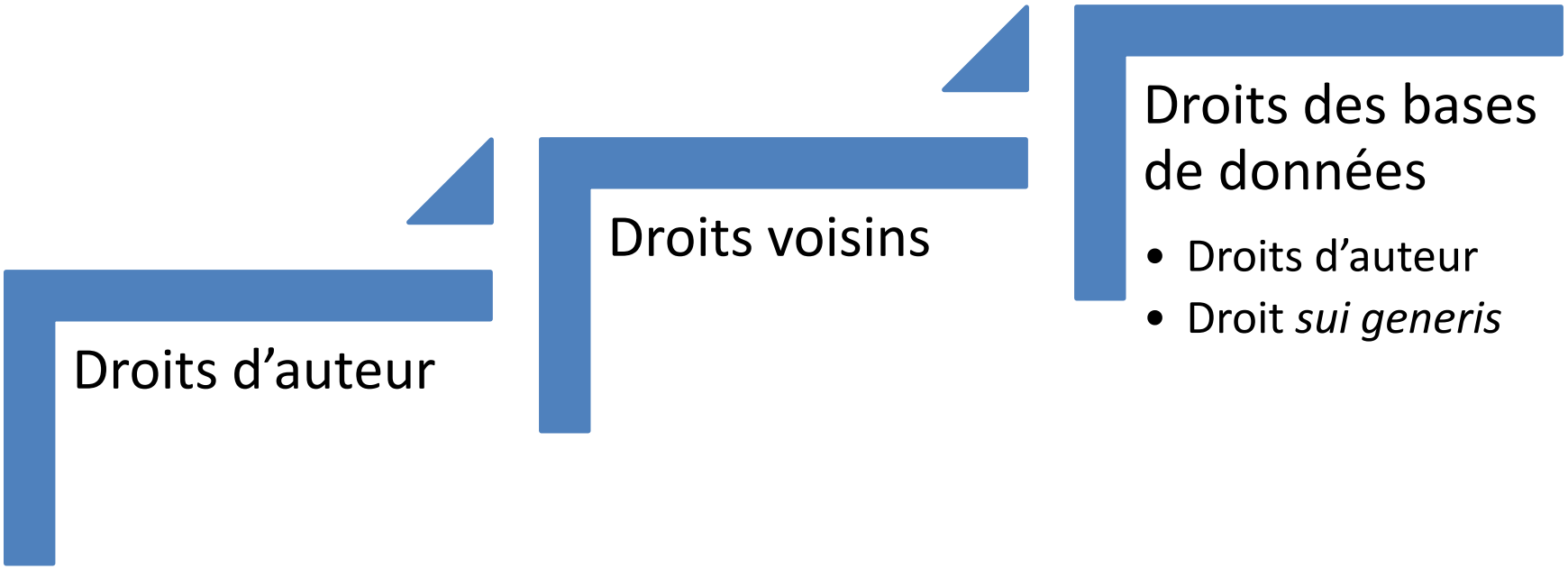
3. Aspects juridiques

1. Le processus du point de vue légal



Inspiré de FutureTDM Practioner Guideline (FTDM, 2010 : 9-10

2. Les types de droits à envisager



4. Quel pays considérer ?

- Les lois de quel pays faut-il considérer dans le cadre d'un projet de TDM ?
  - Le pays où le scientifique réalise sa recherche.
  - Le pays dans lequel les données sont utilisées/copiées/exploitées.
  - Les pays dans lesquels seront publiés les résultats.

3. Le droit européen

- En 2018, diverses négociations au Conseil Européen (*Directive on copyright in the Digital Single Market*)
  - Proposition de la commission : « a mandatory exception allowing research organisation to carry out TDM on content they have lawful access to for scientific research purposes »

3. Le droit européen

- Le 20-21 juin 2018, vote au Parlement Européen
  - Art. 3a proposé comme exception pour la recherche.
- Texte accepté le 12 septembre 2018
  - Art. 3 §1 : “Member States shall provide for an **exception** to the rights provided for in Article 2 of Directive 2001/29/EC, Articles 5(a) and 7(1) of Directive 96/9/EC and Article 11(1) of this Directive for **reproductions and extractions made by research organisations** in order to carry out text and data mining of works or other subject-matter to which they have lawful access for the purposes of scientific research”

4. Aspects scientifiques

Pourquoi fournir ces outils aux chercheurs ?

4.1 Pourquoi le TDM ?

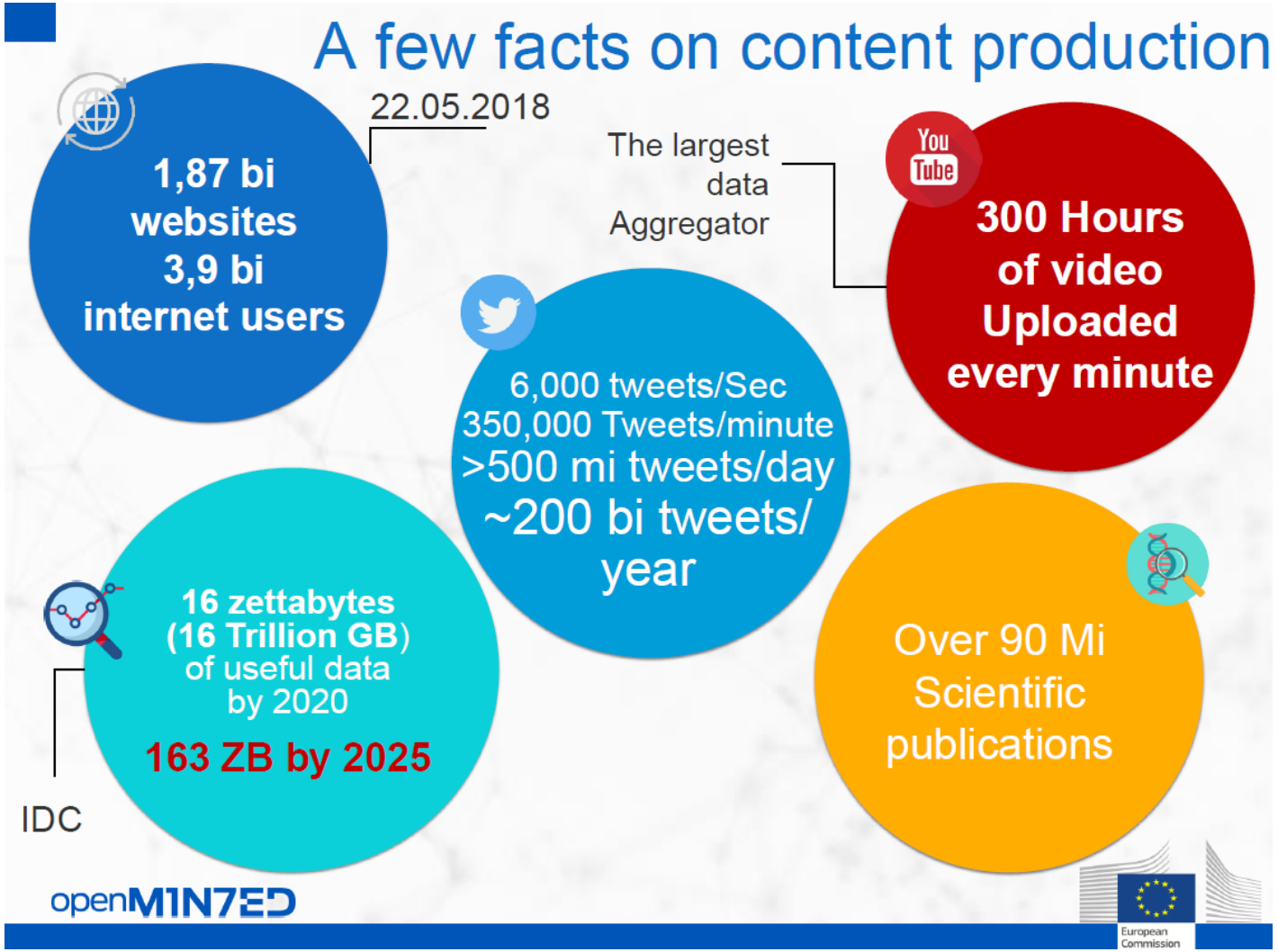


Fig. 5. Data Deluge (Piperidis, 2018))

4.1 Pourquoi le TDM ?

- *Data deluge*
  - 50.000.000 d'articles scientifique (2009)
  - 90.000.000 d'articles scientifique (2018)
  - 426.000 études juste en Chine (2016)
  - 409.000 études juste aux USA (2016)
  - 2.500.000 articles par an en anglais (2015)
  - 1.000.000 d'articles par an sur PubMed

4.2 Pourquoi maintenant ?

- Un (gros) retard en Europe
  - Asie premier centre mondial.
  - USA, doctrine du *fair use* qui permet un usage plus extensif du TDM.
- Une aide au chercheur indispensable
  - « Le TDM favorise la diffusion et la performance de la recherche scientifique » (ADBU)
  - Réduction du temps de recherche dans certain domaine et réduction des couts.

4.3 Pour quelles économies ?

x4	• Facteur de multiplication de la couverture de la connaissance en biologie des systèmes
25%	• Pourcentage du temps de travail économisé pour identifier les études pertinentes en matière de politique publique.
50%	• Amélioration de la productivité en termes de curation de la documentation médicale.
70.000€	• Economie sur une seule base de données suite à l'amélioration de la productivité.

4.4 Pour quoi faire ?

- Exemples de Text Mining
  - Ex. 1 : Analyse du discours
    - Analyse des interventions politiques de différents politiciens afin d'étudier la réparation d'image (Mémoire)
  - Ex. 2 : Analyse linguistique
    - Analyse d'un corpus de textes afin d'étudier les usages sémantico-syntaxiques de l'adverbe maintenant (Mémoire).
  - Ex. 3 : Amélioration curation SV
    - Of Ferulic Acid for Alzheimer's Disease: Combination of Text Mining and Experimental Validation
    - Biomedical Text Mining about Alzheimer's Diseases for Machine Reading Evaluation
  - Ex. 4 : Identifier les tendances (*trends*) de recherche
    - Using Text Mining Techniques to Identify ResearchTrends: A Case Study of Design Research

4.4 Pour quoi faire ?

- Exemples de Citation Mining
  - Ex. 5 : Etudier l'évolution d'un sujet de recherche
    - Mining Google Scholar citation : An Exploratory study.
  - Ex. 6 : Analyser l'écart entre audience réelle et audience visée
    - Integrating Text Mining and Bibliometrics for Research User Profiling.
  - Ex. 7 : Analyser les relations entre auteurs.
    - Mining author relationship in scholarly networks based on tripartite citation analysis.
- Exemples de Data Mining
  - Ex. 7 : Faciliter la FRBRisation
    - Mining bibliographic patterns
    - Mining stars with fp-growth: a case study onbibliographic data
    - A tool for converting from MARC to FRBR

LIÈGE université Library

5. Offre logicielle

5.1 L’offre logicielle

- L’offre logicielle est immense mais comporte 2 grands écueils :
  - Les termes TDM sont des buzzwords
  - Les logiciels sont propres à des tâches/programmes spécifiques
- Plusieurs grammes de logiciels
  - Preprocessing (OpenRefine, Vard2, TextFixer, Porter Stemmer, Lexos)
  - Lexicométrie (IRaMuTeQ, TXM, Modalisa)
  - Data Mining (Weka)
  - Text Mining (Rapidminer)

5.1 L’offre logicielle

- Outils en ligne
  - Projet OpenMindTed
    - Créer une seule plateforme pour rendre visible le contenu en Open Access issu de nombreuses sources.
    - Rendre visible et interopérable de nombreux outils NLP/TDM
    - Faciliter la création et la réalisation de TDM workflow
- Logiciels
  - RapidMiner
  - R (avec quanteda & readtext)

5.2 Exemple : R

- Création de wordcloud
- Création de réseaux lexicaux
  - Lexical Network.txt

LIÈGE université Library

6. Le rôle des bibliothèques

Centraliser les différentes compétences

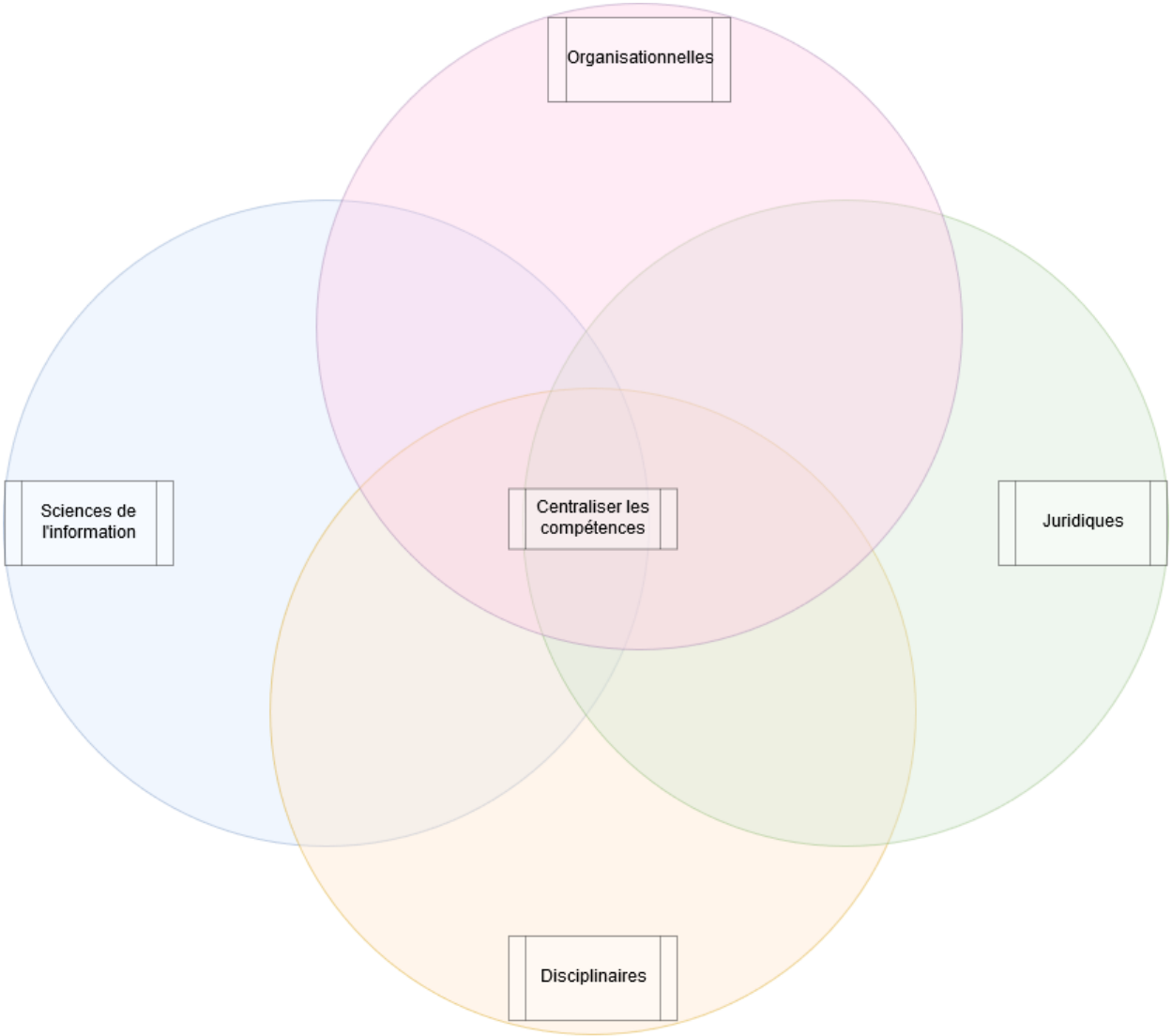
6.1 Un rôle de centralisateur

- L’utilisation des techniques de fouille dans le cadre d’une recherche nécessite :
  - Des ressources matérielles et logicielles
  - Des ressources en sciences de l’information
  - Des ressources disciplinaires
  - Des ressources juridiques
  - D’autres ressources plus spécialisées (informatiques, linguistiques, mathématiques)

6.1 Un rôle centralisateur

- La fouille de données est au cœur des préoccupations informationnelles et bibliothéconomiques :
  - Comment répondre à l'inflation de la littérature scientifique ?
  - Comment automatiser la curation ?

6.1 Un rôle centralisateur



6.2 Pistes et recommandations

- 1. Sensibiliser les chercheurs
  - Promouvoir le TDM
  - Mettre en place des mesures incitatives
  - Proposer des services au sein des bibliothèques
- 2. Développer des expertises
  - Les licences
  - Les outils et les méthodes
  - Le stockage
  - La protection
  - La visualisation

6.2 Pistes et recommandations

- 3. Inscrit le T(D)M dans les licences
  - Exemple fourni parLibLicense Model License

> j. *Text and Data Mining*. Authorized Users may use the Licensed Materials to perform and engage in text and/or data mining activities for academic research, scholarship, and other educational purposes, utilize and share the results of text and/or data mining in their scholarly work, and make the results available for use by others, so long as the purpose is not to create a product for use by third parties that would substitute for the Licensed Materials. Licensor will cooperate with Licensee and Authorized Users as reasonably necessary in making the Licensed Materials available in a manner and form most useful to the Authorized User. If Licensee or Authorized Users request the Licensor to deliver or otherwise prepare copies of the Licensed Materials for text and data mining purposes, any fees charged by Licensor shall be solely for preparing and delivering such copies on a time and materials basis.

- 4. Développer l'infrastructure

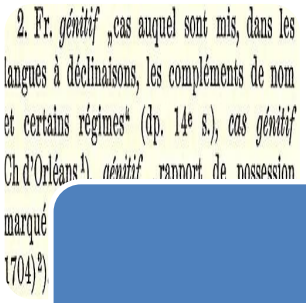
7. Conclusion

. Conclusion

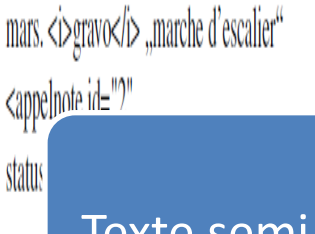
- La fouille de données sert médiatement ou immédiatement la recherche scientifique.
- Mais cela nécessite des compétences variées.
  - Compétences techniques
  - Compétences juridiques
  - Compétences disciplinaires
- Il faut donc un facilitateur.
  - Rôle des bibliothèques universitaires ?

Annexes

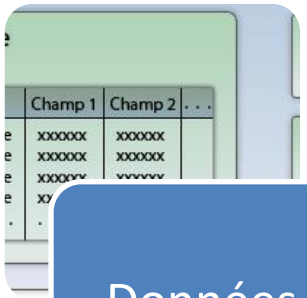
Annexes. Aller plus loin : un espace vectoriel


2. Fr. *génitif* „cas auquel sont mis, dans les langues à déclinaisons, les compléments de nom et certains régimes“ (dp. 14e s.), *cas génitif* Ch d'Orléans 1) *génitif* rapport de possession marqué 1704) 2)

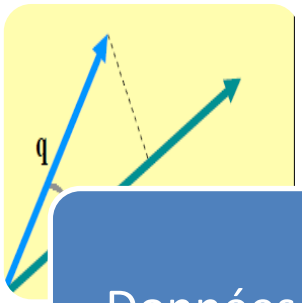
Texte brut


mars. gravoc „marche d'escalier“
<appelnote id="2"
status

Texte semi structuré



Données structurées



Données vectorielles

- Réduire le nombre de mots (*preprocessing*)
  - Faible valeur sémantique (*stop words*, etc.)
  - La loi de Zipf [$f(n) = \frac{k}{n}$]
- Créer un tableau de données
  - Une ligne par texte
  - Une colonne par *mot/token*
  - Une cellule pour la fréquence
- Créer des vecteurs
  - Chaque document est représenté par un vecteur
  - Les coordonnées du vecteur sont les mots/tokens.

$$D_j := \begin{bmatrix} f_{1j} \\ f_{2j} \\ f_{n_w j} \end{bmatrix}$$

- D_j représente le document
- f_{1j} représente la fréquence du mot W_i
- n_w représente le nombre total de mots
- Il s'agit simplement d'un « *bag of words* »

$$\mathbf{D} \triangleq \overbrace{\begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1n_d} \\ f_{21} & f_{22} & \cdots & f_{2n_d} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n_w 1} & f_{n_w 2} & \cdots & f_{n_w n_d} \end{bmatrix}}^{\text{documents}} \Bigg\} \text{words}$$

- Quel est l'intérêt de ce type de représentation ?
  - Les documents et les recherches sont représentés par des vecteurs.
  - Il est alors possible de retrouver un document à l'aide de la notion de similarité.
  - Autrement dit, il est possible d'appliquer des algorithmes plus performant puisque nous n'avons plus que des données numériques.

- Poids des mots (*term weighting*)
  - Définir la probabilité qu'un mot apparaisse dans les documents pour définir son poids.
  - Ou il s'agit de calculer l'inverse de la fréquence du mot dans le document (*idf score*)
  - Il devient aussi possible de calculer l'importance d'un mot dans un seul document (proportionalisé) (*fr score*)
  - Calculer le *tf-idf* score (multiplication des deux)

– « Co-occurrence analysis » réalisée avec R