

TextBI : Une plateforme générique pour la visualisation interactive de données multidimensionnelles issues des réseaux sociaux

Maxime Masson, Christian Sallaberry, Marie-Noelle Bessagnet,

Philippe Roose, Annig Le Parc Lacayrelle,

Rodrigo Agerri

Journées du GdR CNRS MAGIS

Atelier AR 7 « Humanités Numériques Spatialisées »

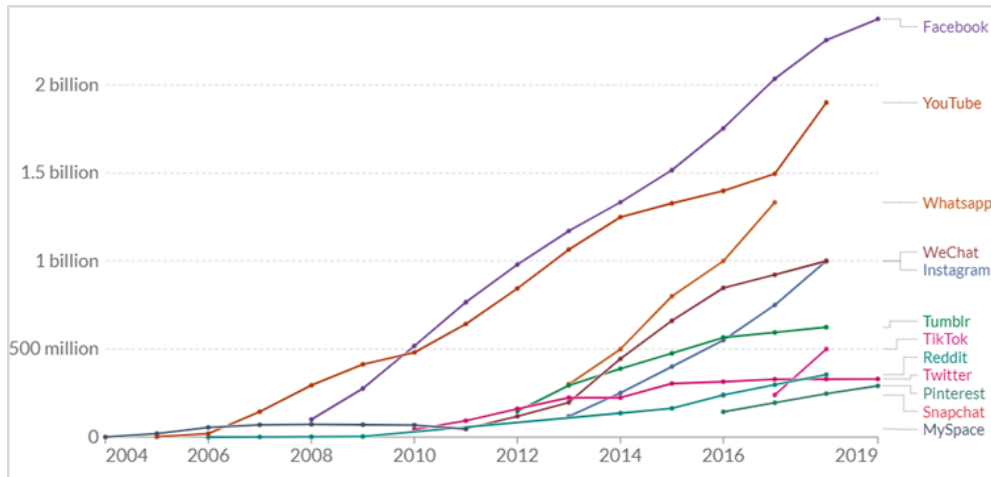
24 novembre 2023

Bordeaux, France

Durée: **15 minutes**

Contexte — Contenu généré par les utilisateurs (UGC)

- **Croissance significative** des sources de données disponibles dans de **nombreux domaines**
 - De plus en plus massives et diversifiées
- **Contenu Généré par les Utilisateurs (UGC)**
 - *Réseaux sociaux, blogs, sites d'avis, etc.*



Évolution du volume de personnes utilisant les principaux réseaux sociaux, de 2004 à 2019

Source: *Our World In Data*

Motivations — Besoins des professionnels du tourisme

Professionnels du tourisme

Offices de tourisme, mairies, organisations de marketing des destinations (DMOs), etc.

- Assistance dans le **processus de prise de décision** et la planification des infrastructures
- Compréhension des **exigences**, des **pratiques** et des **attentes** des visiteurs (adaptation)

Identifiés en collaboration avec la
communauté d'agglomération
du Pays Basque (Bayonne)

Quels sont des lieux et
activités **méconnus**
mais agréables qui
devraient être mis en
avant ?

Quelles sont les
pratiques touristiques
les plus associées à
chaque municipalité ?

Quelles plages
populaires ont des
infrastructures
inadéquates ?

Quels événements sont
les plus populaires
pendant la période
estivale ?

Nous faisons l'hypothèse que les **réseaux sociaux** permettraient de répondre à **ces besoins métiers**

Motivations — Besoins des touristes

Touristes (*visiteurs ou habitants*)

Recommandation

- Activités Touristiques
- Point d'Intérêt (POI)
- Itinéraires



Mise en relation d'utilisateurs

- Intérêts Communs
- Intérêts Divergent



Défis — Traitement des publications sur les réseaux sociaux

Les réseaux sociaux présentent **de nombreux défis**...

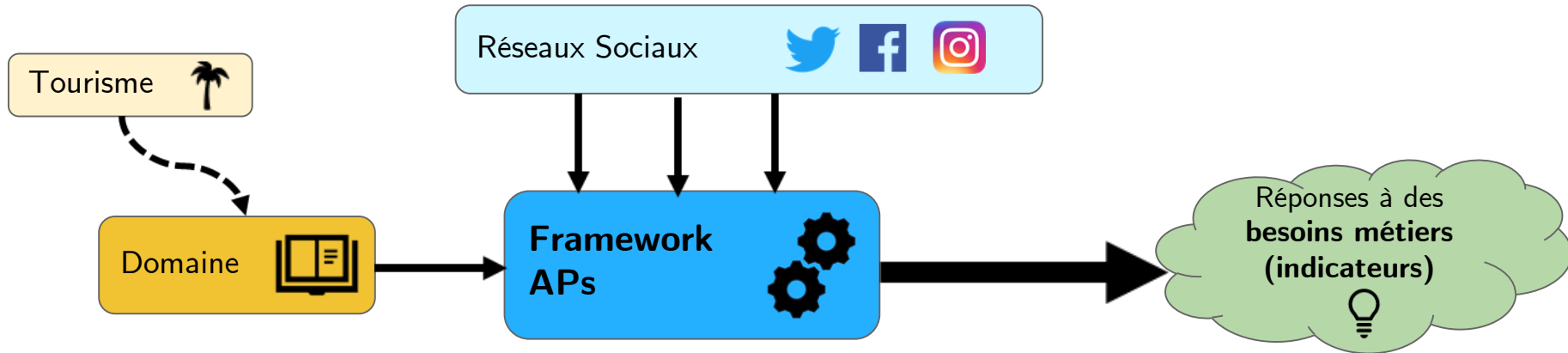
- ❑ Textes courts
- ❑ Textes non structurés ou mal structurés
- ❑ Langage informel
- ❑ Présence d'emojis, de hashtags, d'URLs
- ❑ Mal orthographiés et mal ponctués
- ❑ Multilinguisme
- ❑ Très grand nombre de publications
- ❑ Beaucoup de bruit

... mais ont aussi des **avantages significatifs** :

- ❑ Données à jour, en temps réel
- ❑ Grand volume de données
- ❑ Richesse, couvrant divers sujets
- ❑ Peu cher, contrairement aux données commerciales.

Framework APs — Vue d'ensemble

- Un nouveau **cadre de travail (framework)**
 - **APs** : *Augmented Proxemics services*
 - **Objectif** : guider l'analyse des données des réseaux sociaux en utilisant la théorie de la proxémique
- **Entièrement générique** mais expérimentation dans le **domaine du tourisme** et la **région du Pays Basque**.



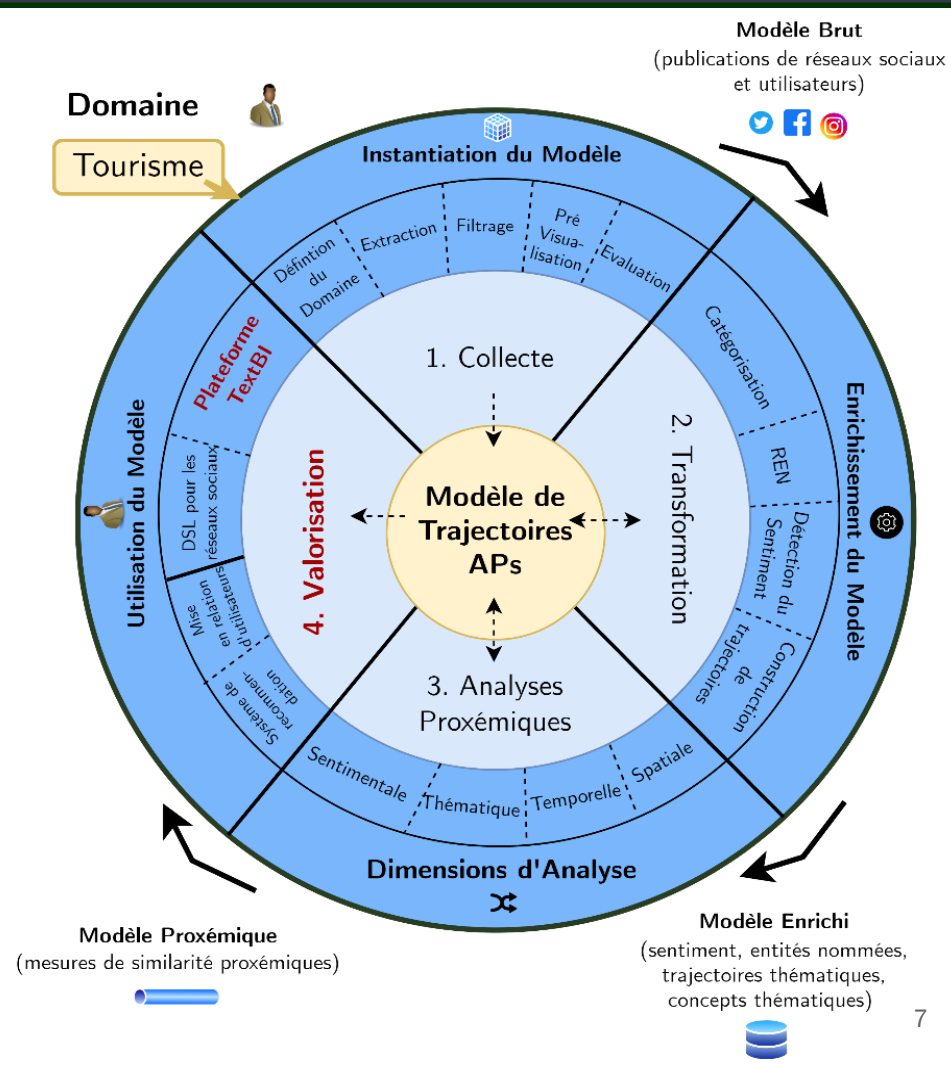
Framework APs — Cycle de vie

- 4 étapes

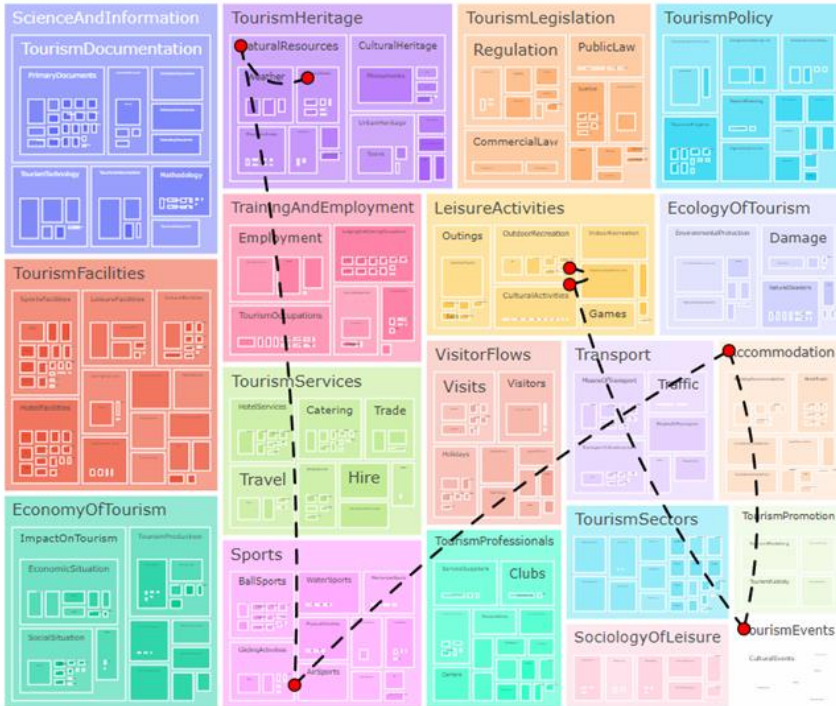
- Collecte
- Transformation
- Analyses Proxémiques
- Valorisation

- Générique

- Sur la **source** et sur le **domaine**
- **Multilingue**
- Basé sur les **dimensions d'analyse**
 - Spatiale (S)
 - Temporelle (T)
 - Thématique (TH)



Modélisation du domaine tourisme



- Projection des trajectoires des utilisateurs sur une **carte thématique du tourisme**
 - *Ontologie, Thésaurus, Dictionnaire*
 - Ubiquité
- Basé sur le **Thésaurus Tourisme et Loisirs de l'Organisation Mondiale du Tourisme (OMT)**
 - **Terminologie multilingue** étendue (FR, EN, ES)
 - Plus de 2 500 concepts thématiques
 - 20 branches principales

1. Collecte

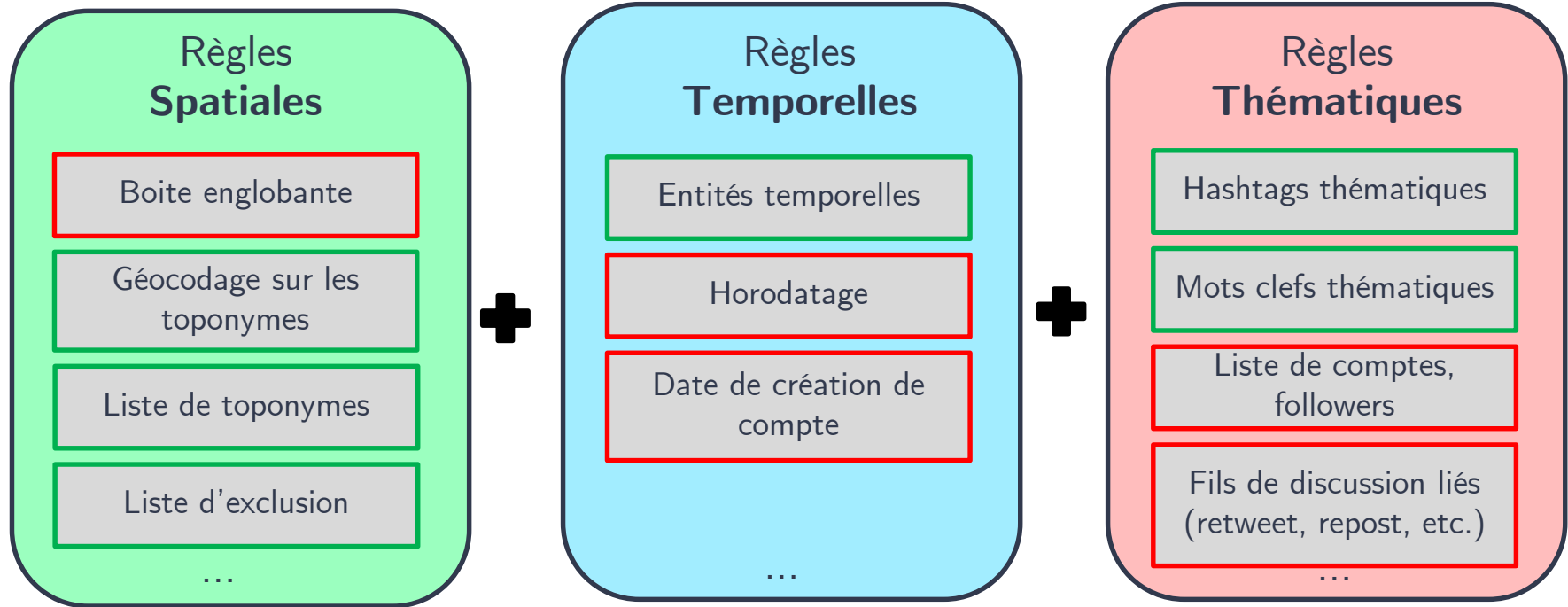
Collecte de données — Techniques Existantes



Basé sur le contenu



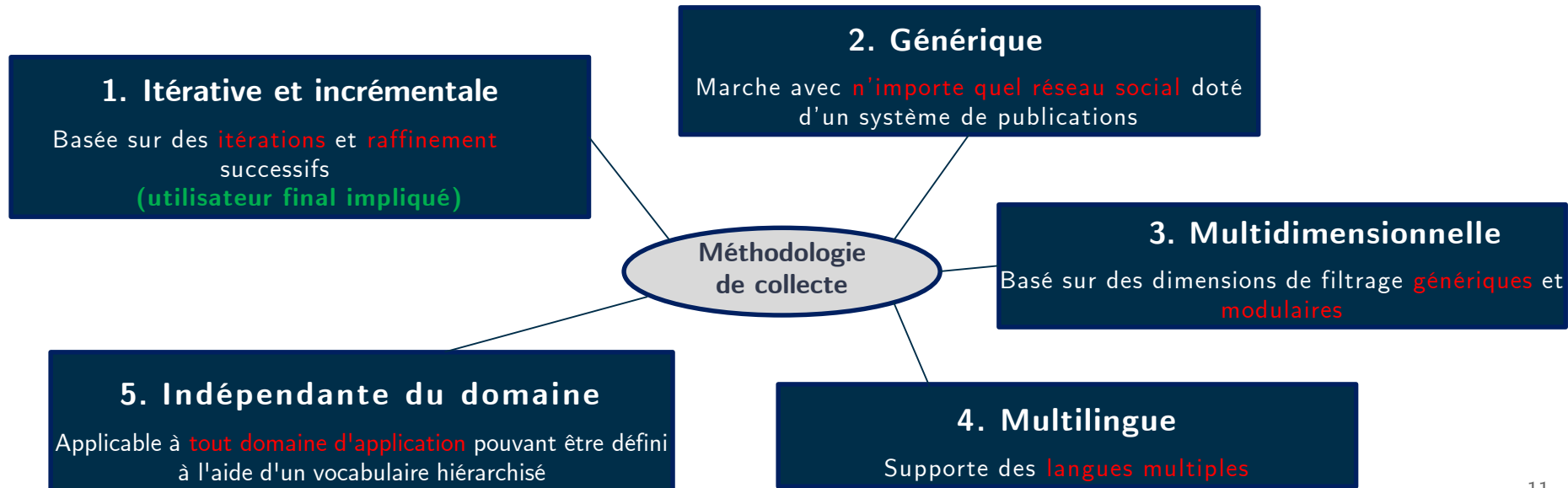
Basé sur les métadonnées



Pas d'approche de collecte **générique de haut niveau**. Principalement des implémentations **ad hoc**.¹⁰

Méthodologie de collecte générique — Caractéristiques

- Conception d'une **méthodologie itérative** pour **construire un jeu de données** sur les réseaux sociaux autour d'un **thème**, d'une **spatialité** et d'une **temporalité** donnés.



Méthodologie de collecte générique — Dimensions

Spatiale

Empreinte
Spatiale

Exemple

*Le Pays Basque
L'Espagne*

*Boîte englobante
Toponymes*

Temporelle

Période
temporelle

Exemple

*L'été 2019
Aout 2020*

*Date et heure
Saison, jour de la semaine
etc.*

Thématique

Sémantique des
données

Exemple

*Le Tourisme, La Mode,
Les Politiques Publiques*

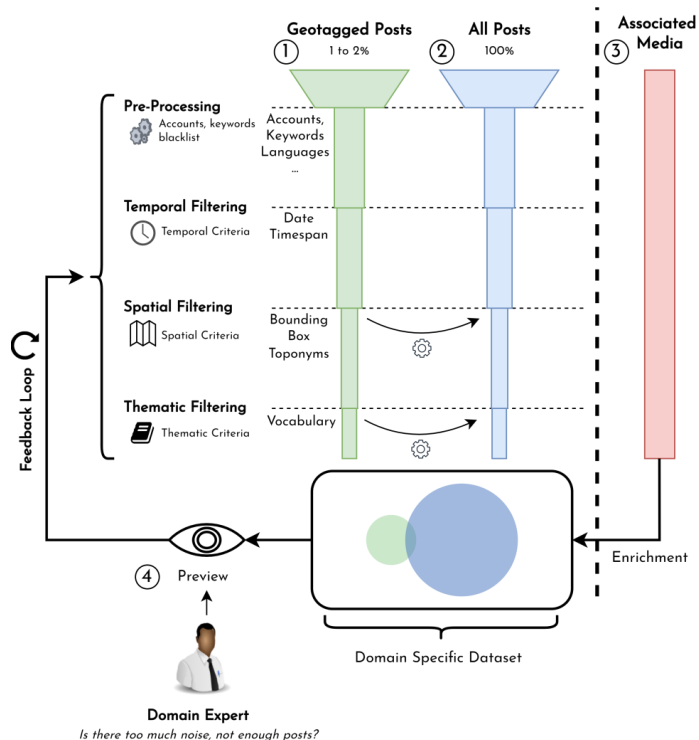
*Dictionnaire
Thésaurus
Ontologie*

Collaboration avec
l'utilisateur final



Définition du jeu de données

Méthodologie de collecte générique — Processus



- **2 flux de données**
 - Publications **géolocalisées** (1 à 2%)
 - Autres publications (> 98%)
- **4 étapes successives de filtrage**
 - Pré-filtrage
 - Filtrage temporel
 - Filtrage spatial
 - Filtrage thématique
- **Mécanisme de boucle de feedback**
 - Aperçu et évaluation du jeu de données
 - Feedback sur le filtrage

Méthodologie de collecte générique

Expérimentation



Côte Basque Française



Été 2019

August 2019							September 2019						
Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat
				1	2	3	1	2	3	4	5	6	7
4	5	6	7	8	9	10	8	9	10	11	12	13	14
11	12	13	14	15	16	17	15	16	17	18	19	20	21
18	19	20	21	22	23	24	22	23	24	25	26	27	28
25	26	27	28	29	30	31	29	30					

www.a-printable-calendar.com

Thésaurus du Tourisme
et des loisirs de l'OMT



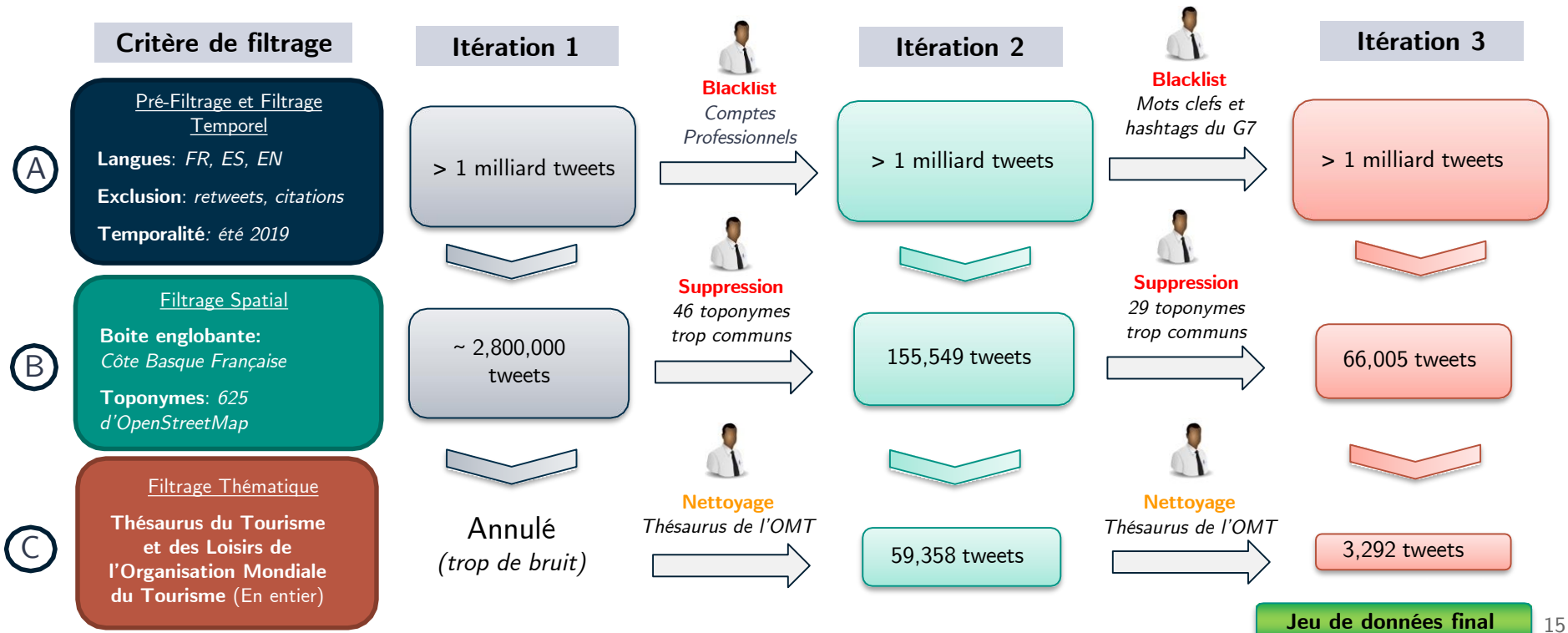
UNWTO

World Tourism Organization

Source: <https://www.e-unwto.org/doi/book/10.18111/9789284404551>

Méthodologie de collecte générique

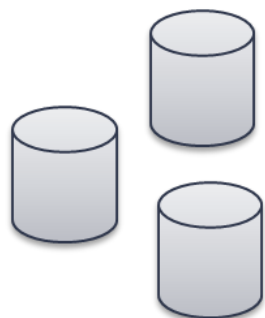
Expérimentation - Processus



2. Transformation

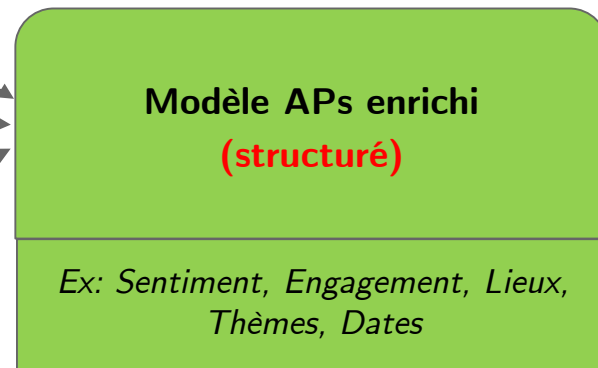
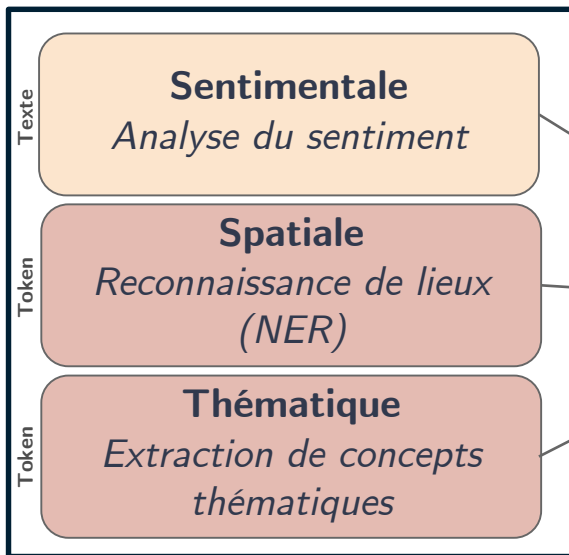
Transformation — Vue d'ensemble

Batterie de modules de TAL (NLP) Machine Learning et basés sur des Règles



Corpus brute issu de réseaux sociaux – collecté durant la phrase précédente
(non structuré)

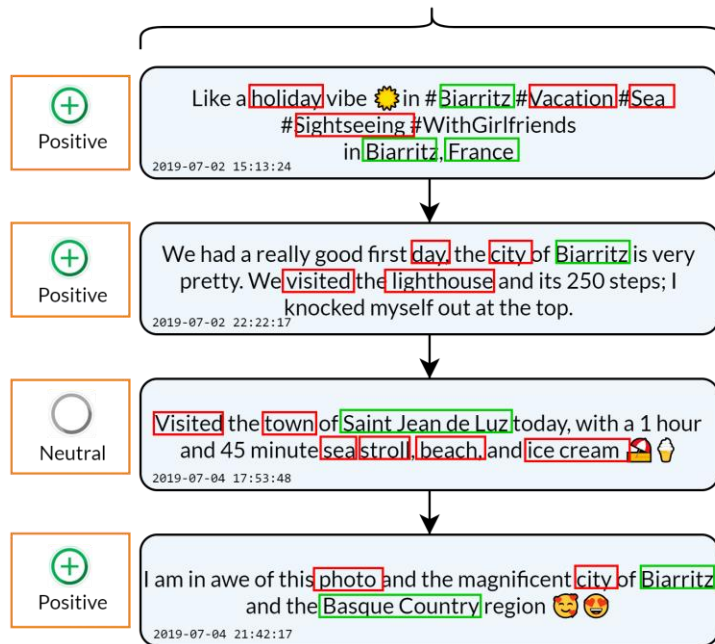
Collection de **publications** sur les réseaux sociaux et leurs **métadonnées associées**






Engagement, Utilisateurs, Géolocalisations

Transformation — Tâches

Tweet Trajectory of a Tourist

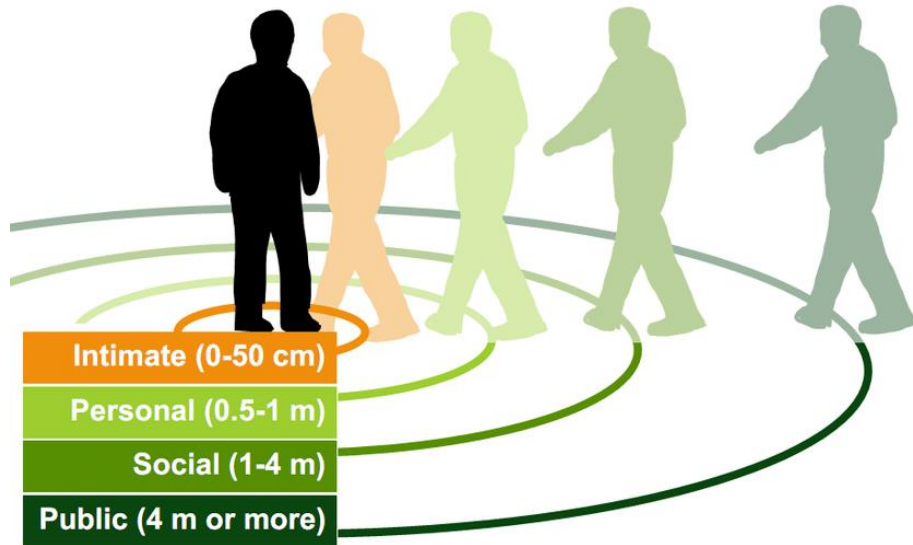


-  Sentiment Annotations (text)
-  Touristic Concept Annotations (token)
-  Locations Annotations (token)

3. Analyses Proxémiques

Proxémique — Introduction

La science qui étudie l'effet de l'espace et des distances sur les interactions et les comportements (Hall, 1966).



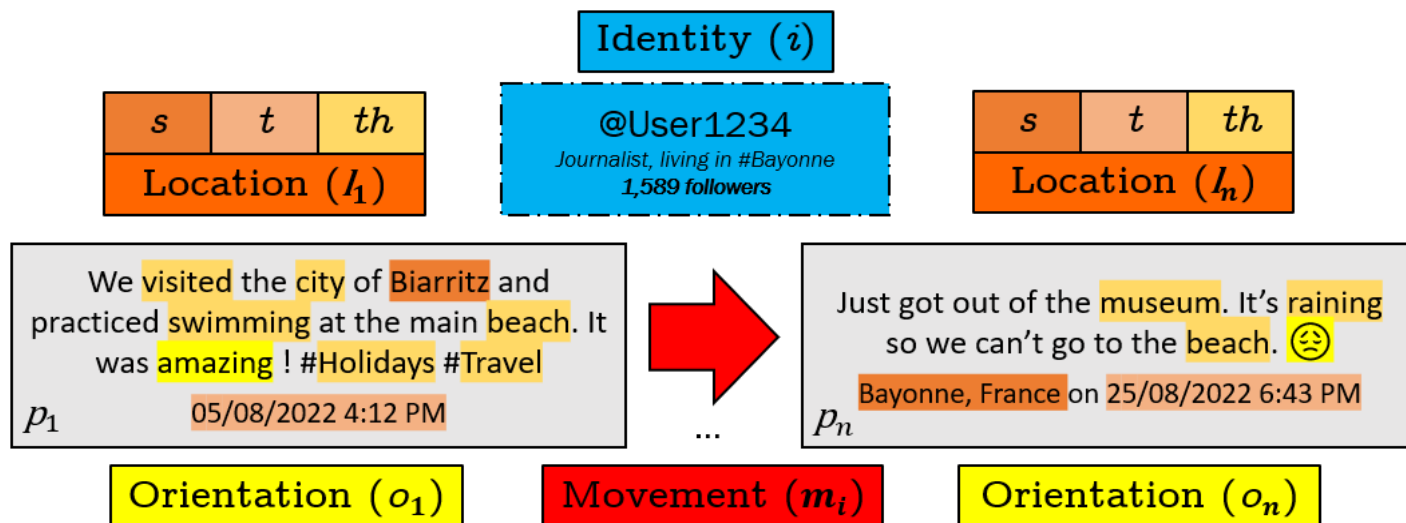
Source: Marquardt, Nicolai & Greenberg, Saul. (2012).

Informing the Design of Proxemic Interactions. IEEE Pervasive Computing

- **Dimensions DILMO (Greenberg, 2011) :**
 - Distance
 - Identité
 - Localisation
 - Mouvement
 - Orientation
- **Niveau d'analyse**
 - Individuel
 - Groupe
- Notion de « **centralité** »
- Comment cette **théorie physique** peut-elle être **adaptée pour modéliser les interactions sur les réseaux sociaux** de manière **générique** ?

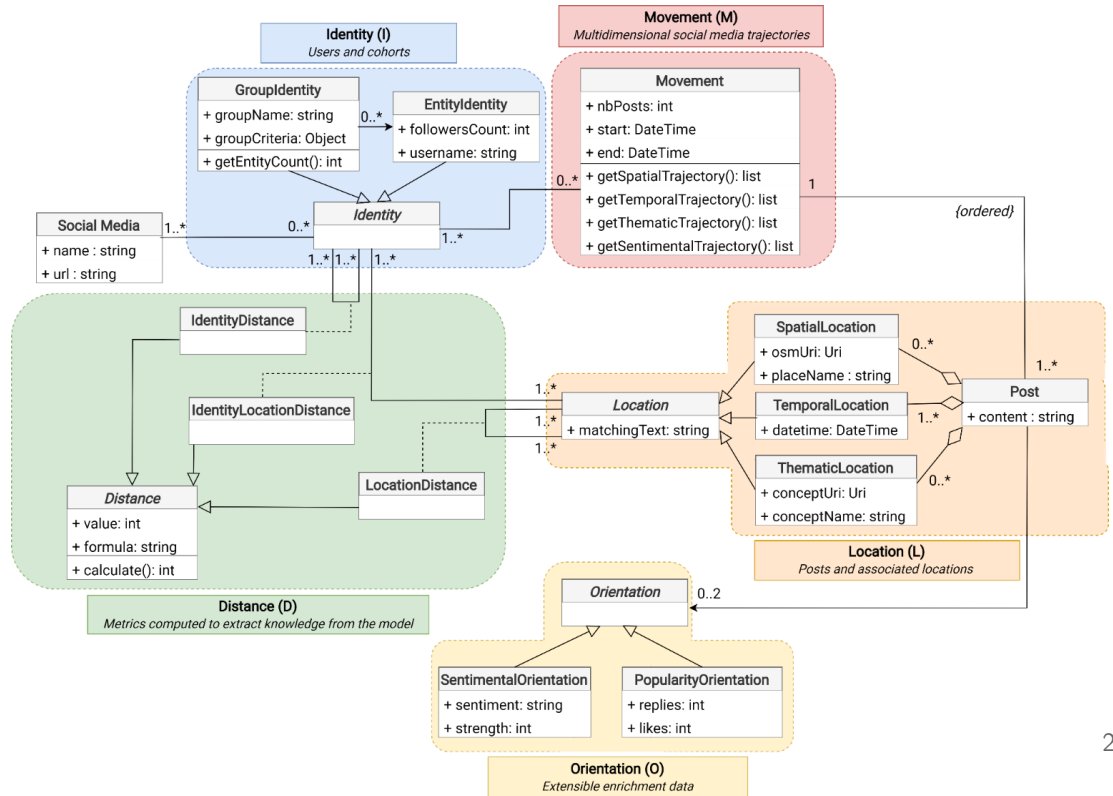
Proxémique — Redéfinition

- Redéfinition de la proxémiques pour les réseaux sociaux

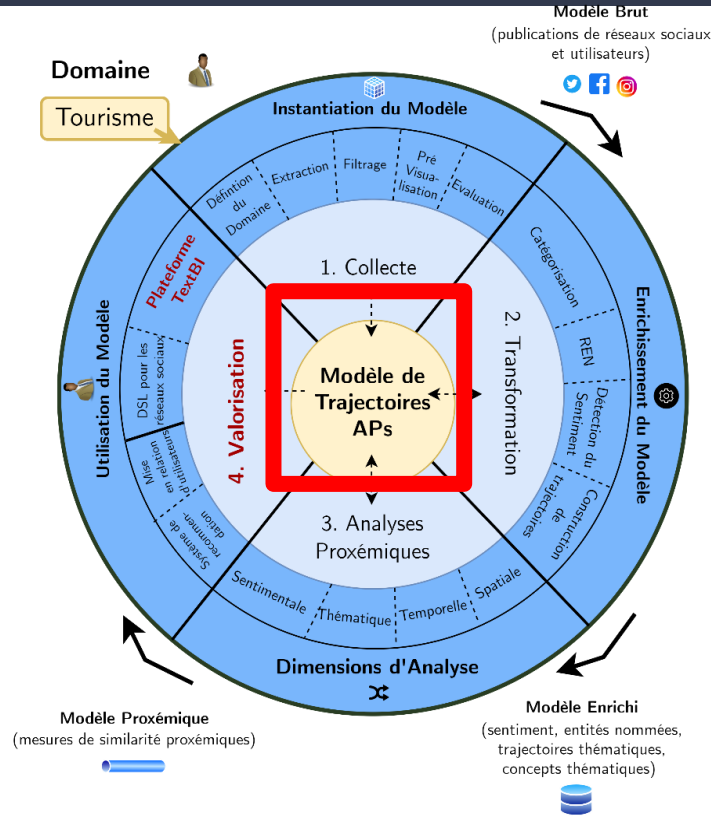


Modèle de Trajectoires Proxémique APs — Diagramme de classe

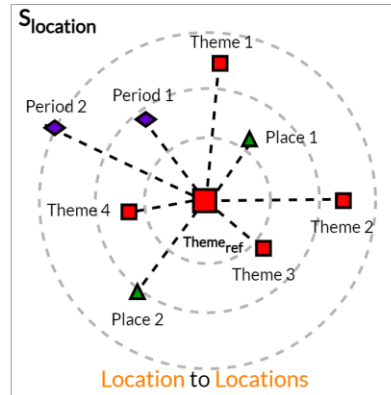
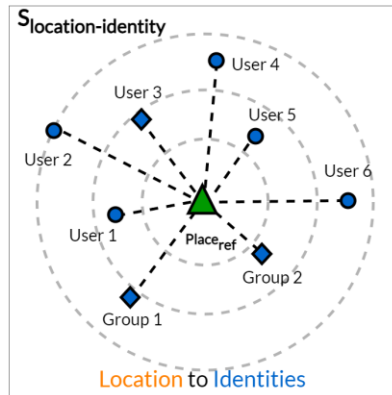
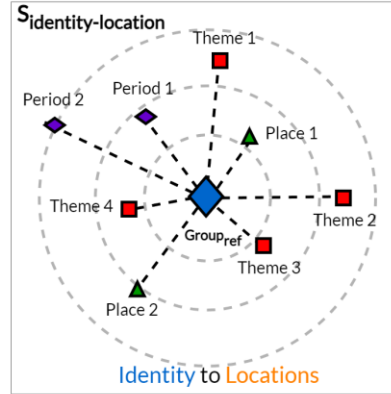
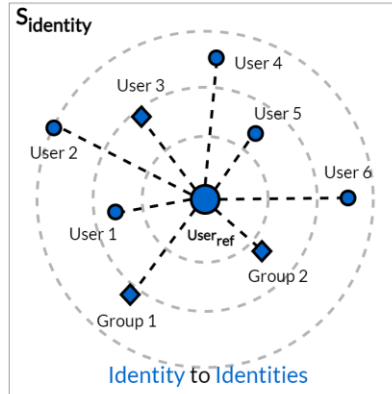
- Modèle de trajectoire générique
- Adaptable à tout domaine d'application
 - **Générique**
 - **Extensible**
- Incorpore les 5 dimensions de la proxémique
 - **Distance**
 - **Identité**
 - **Localisation**
 - **Mouvement**
 - **Orientation**



Modèle de Trajectoires Proxémique APs - Rôle



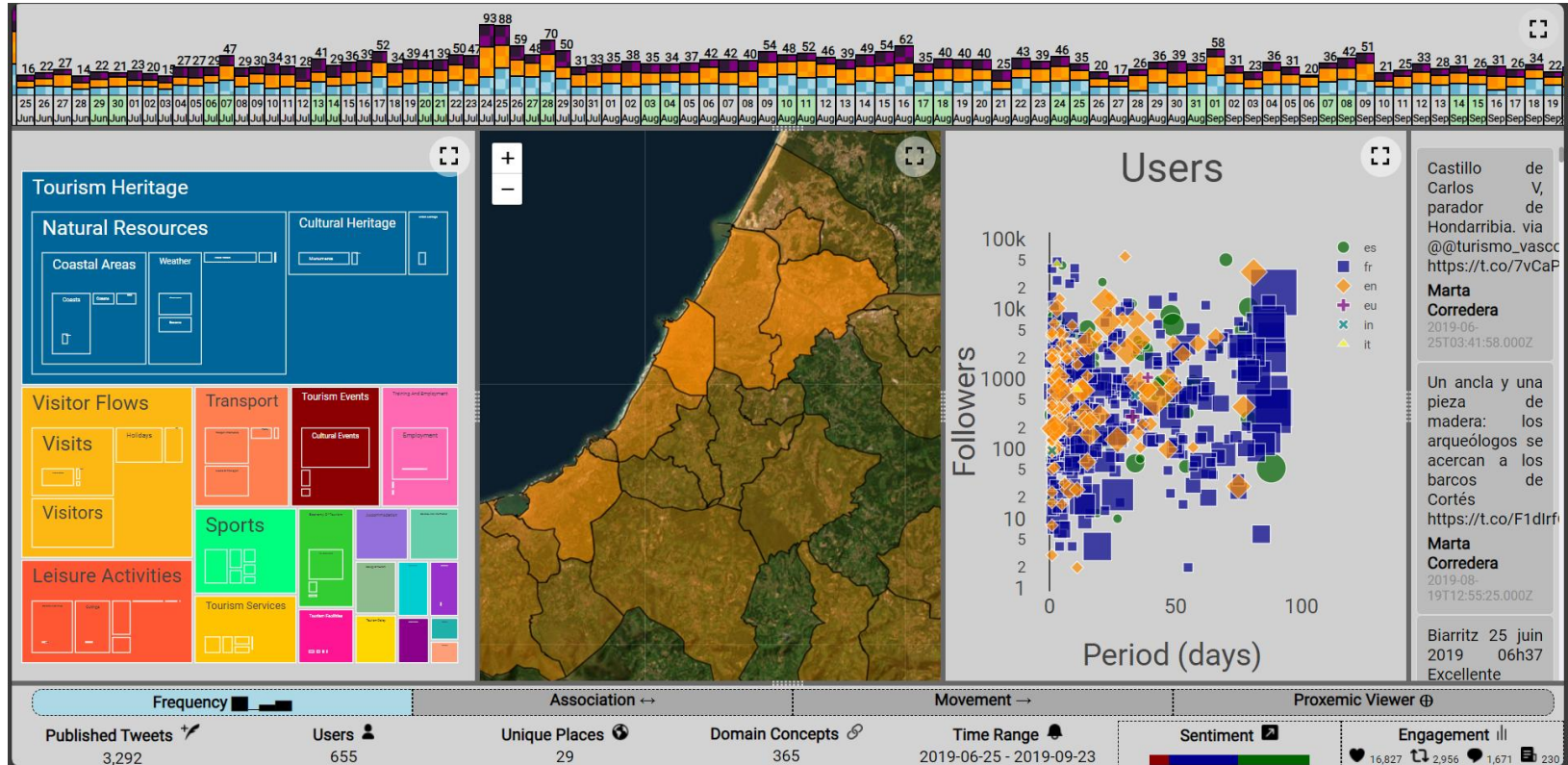
Boite à outil pour évaluer la similarité proxémiques



4. Valorisation

Valorisation — Plateforme TextBI

- Une **plateforme générique** (de type tableau de bord) permettant de **visualiser des corpus de réseaux sociaux annotés**.

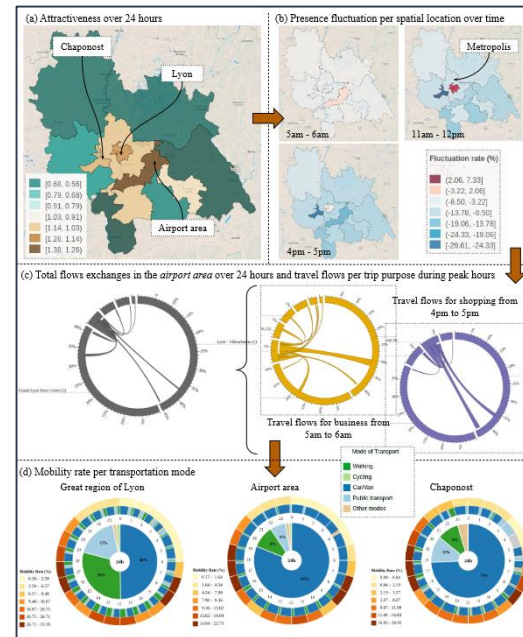
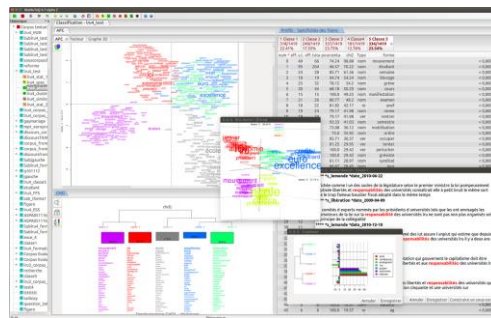


TextBI - Utilisateurs

- 2 catégories d'utilisateur
 - **Décideurs du domaine** (*non-informaticien*) :
 - Effectuer des analyses sur leur domaine, aider à la prise de décision
 - **Ex:** *office de tourisme cherchent à identifier les activités touristiques dominantes dans une ville*
 - **Chercheurs** (*informaticien*)
 - Evaluer des processus et modèles de TAL
 - **Ex:** *chercheurs en TAL cherche à visualiser la distribution des entités nommées dans un corpus de données issue des réseaux sociaux*
- **Besoins communs**
 - *Distribution, Fréquence, Association, Enchaînement, Etc.*

TextBI - Inspirations

- Outils « Géomatique »
 - QGIS, ArcGIS, Travaux d'Aline Ménin et de Cécile Saint-Marc
- Outils « Business Intelligence (BI) »
 - Qlik, Tableau, Power BI ...
- Outils « Traitement Automatique du Langage (TAL) »
 - IRaMuTeQ, Voyant, VOSviewer, SentimentViz



TextBI - Inspirations

	Outils GIS	Outils BI	Outils TAL
Exemple	QGIS, ArcGIS	Power BI, Tableau Public, QlikView	SpaCy, GATE, TextRazor, IRaMuTeQ, VOSViewer, SentimentViz
Points Forts	<ul style="list-style-type: none">Analyse spatiale pousséeGranularités spatiales multiples	<ul style="list-style-type: none">Exploration de données multidimensionnellesSimple d'utilisationInteractivité inter et intra visuelFiltrages combinésSynchronisation des visuels	<ul style="list-style-type: none">Statistiques avancées (cooccurrence, fréquences, etc.)Gestion native du texteVisuels plus adaptés au texte : graphes sémantiques, nuages de mot, etc.
Limitations	<ul style="list-style-type: none">Difficile de visualiser des données sans composante spatialePeuvent être complexe pour des utilisateurs néophytes (non géographes)	<ul style="list-style-type: none">Nécessite des données structurées (e.g., base de données relationnelle)Principalement pour des données numériquesDifficultés à visualiser des données séquentielles	<ul style="list-style-type: none">Complexe, vise principalement des utilisateurs spécialistes du TALAnalyse au niveau des mots ou focus sur une seule dimensions (ex: sentiment)

Démonstration de TextBI

Bilan

Bilan et Perspectives Futures

- Prise en main de TextBI par des **utilisateurs finaux** (office de tourisme)
- Expérimentation additionnelle sur ...
 - **Un autre domaine que le tourisme** pour **démontrer la généricité de la plateforme**
 - **Exemple** : *Politiques Publiques*
 - **Une autre source de données**
 - Massive (test de la scalabilité de la plateforme)
 - Autre réseau social :
 - **Exemple** : site d'avis sur les villes
- Ajout du support pour des flux de **données temps-réel** (suivi d'une élection, d'un évènement sportif, etc.)
- **Amélioration interface**
 - Mode Proxémique
 - Sélection de la granularité
 - Chargement des données
- **Mise à disposition de la plateforme à la communauté (open source)**

Publications

- **[2022]** Masson, M. **Services augmentés pour le tourisme intelligent et l'analyse des pratiques.** In *Forum Jeunes Chercheuses Jeunes Chercheurs*, INFORSID 2022, Dijon, France (p. 35)
- **[2022]** Masson, M., Sallaberry, C., Agerri, R., Bessagnet, M. N., Roose, P., & Le Parc Lacayrelle, A. (2022, October). **A domain independent method for thematic dataset building from social media: the case of tourism on Twitter.** In *International Conference on Web Information Systems Engineering – WISE 2022*, Biarritz, France (pp. 11-20). Cham: Springer International Publishing.
- **[2023]** Masson, M., Roose, P., Sallaberry, C., Agerri, R., Bessagnet, M. N., & Lacayrelle, A. L. P. (2023, April). **APs: A Proxemic Framework for Social Media Interactions Modeling and Analysis.** In *International Symposium on Intelligent Data Analysis – IDA 2023*, Louvain-La-Neuve, Belgium (pp. 287-299). Cham: Springer Nature Switzerland.
- **[2023]** Masson, M., Abdelhedi, S., Sallaberry, C., Agerri, R., Bessagnet, M. N., Lacayrelle, A., & Roose, P. **Visualisation interactive de trajectoires d'activités touristiques.** *Workshop Exploration des traces dans un monde du tout numérique : enjeux et perspectives – INFORSID 2023*, La Rochelle, France, 12.

Challenges Géodata – GéoDataDays 2023

Découvrez les lauréats des Challenges Geodata 2023 :

1er : Visualisation de données issues des réseaux sociaux : une plateforme de type Business intelligence

Maxime MASSON, Laboratoire LIUPPA

2ème + prix du public : HedgeTools : une boîte à outils pour extraire et caractériser les haies en milieu

agricole - Gabriel MARQUES, Dynafor

3ème : AEROLAB SPACE - Lilian JOLY, Université de Reims Champagne-Ardenne

4ème : Cartographie scénarisée et sensibilisation aux risques naturels - Thomas CANDELA, RisCrisis / LAGAM

5ème : Une boîte à outils Python pour utiliser un eye-tracker avec des cartes - Laura WENCLICK, LASTIG, IGN, Univ Gustave Eiffel, ENSG

6ème : MapDraw – un outil pour annoter des cartes interactives et multi-échelles - Justin BERLI, LASTIG, IGN, Univ Gustave Eiffel, ENSG

7ème : Prévion de l'aléa incendie par télédétection optique/thermique haute résolution - Victor PENOT, CESBIO

8ème : Etude de trajectoires de déplacement dans un contexte spatiale et temporelle en relation avec une expertise du domaine - Rouaa WANNOUS, IUT de l'Université de la Rochelle

Merci pour votre attention !

Des questions ?

maxime.masson@univ-pau.fr