APPEL À PROJETS 2022 – DONNÉES & OPEN DATA

FÉDÉRATION DE RECHERCHE AGORANTIC
« CULTURE, PATRIMOINES, SOCIÉTÉS NUMÉRIQUES »

Titre	Mise en correspondance de deux corpus de données : appréhension par l'interdisciplinarité des différentes formes textuelles de l'objet « spectacle »
Acronyme	OduS2Apidae
Nom du/des porteur(s)	Juan-Manuel Torres-Moreno, Pierre Jourlin, Eric Triquet
Laboratoires associés	LIA, CNE
Budget demandé	6.000€
Résumé *Max. 1 000 caractères espaces compris*	Cette proposition s'inscrit dans le prolongement du projet OduS et apporte à ce dispositif un module complémentaire destiné à la recherche. OduS permet de créer des données de qualité scientifique dans le domaine du spectacle vivant. OduS2Apidae est un module qui permet de créer des correspondances entre les données d'OduS et les données de la plateforme de mutualisation de l'information touristique Apidae. La mise en correspondance de ces deux corpus aux propriétés différentes permet de réaliser des recherches en informatique (entraînement et évaluation) et en SHS (analyse de la qualité des données de la plateforme Apidae). Cet appel à projet nous permettra de réaliser l'outil OduS2Apidae et de réaliser de premiers travaux interdisciplinaires à partir des corpus obtenus.

1. Contexte, positionnement, objectif(s)

Contexte et origine

Le projet OduS2Apidae s'inscrit dans un projet de recherche et d'innovation plus large. Depuis 2016, nous travaillons sur une plateforme de production de données destinée à obtenir des données de qualité scientifique (exhaustives, homogènes et structurées) pour différents domaines d'activités socio-économiques. Les données qui en sont issues sont destinées à la recherche et au développement socio-économique des territoires. Cette plateforme a été conçue à partir des travaux doctoraux d'Eloi Flesch et le développement du dispositif est soutenu par l'InSHS, la Satt Sud-Est et le ministère de la culture.

Pour sa première expérimentation et à des fins de preuve de concept, le dispositif OduS sera dédié aux données du spectacle vivant et le projet porté sur le territoire des Hautes-Alpes par le Théâtre Du Briançonnais et il est soutenu par les collectivités locales, le conseil départemental, la région Sud et le programme européen *Leader*. Sur ce volet territorial, la plateforme permettra de mutualiser les données culturelles de l'ensemble des acteurs culturels pour les diffuser en open data sur une pluralité de supports (sites web et applications mobiles, plateformes, agenda culturels, blogs, etc.) En outre, sur le volet recherche, le dispositif permet d'enquêter sur les pratiquants du domaine à partir d'une méthodologie intedisciplinaire *ad-hoc*[1].

[1] Ce travail d'enquête financé via un contrat de collaboration de recherche entre le CNE et le Théâtre Du Briançonnais est prévu d'avril à décembre 2023 avec une équipe de quatre chercheurs et un ingénieur d'études.

Objectifs

OduS2Apidae va permettre de croiser les données des deux plateformes : de mettre en correspondance les données d'Apidae (plateforme de mutualisation de l'information touristique) produites par les offices du tourisme avec celles produites scientifiquement avec OduS. C'est en mettant en comparaison la qualité des données de ces deux dispositifs traitant d'un même objet[2] que nous allons pouvoir ouvrir de nouvelles perspectives de recherche d'une part en SHS (étude des structures informelles de textes de présentation de spectacles, d'autre part en informatique (étude des structures informelles de textes de présentation de spectacles). Voir section : *5. Interdisciplinarité*

2. Questionnement scientifique en rapport avec l'intitulé de l'appel

L'association de ces deux corpus de données par des correspondances entre leurs instances d'objets ouvre de nouvelles perspectives de recherche en informatique et SHS en comparant deux sources de données d'un même domaine. Le premier de ces corpus (Apidae) est produit dans un cadre peu formalisé,invitant ses utilisateurs à s'affranchir de toute structuration et sémantiques communes (ceci aboutit à un corpus à très forte hétérogénéité sur les plans de la structuration et de la qualité des contenus - voir le projet DataTour qui s'intéresse aux raisons de cette hétérogénéité). Le second (OduS) produit des données structurées et homogènes, sémantiquement peu ambigüe ; celui-ci constitue notre corpus de référence pour étudier les données hétérogènes d'Apidae.

Dans le cadre de ce projet nous aborderons deux questionnements scientifiques et leur questionnement commun est celui des formes et de leurs reconnaissances.

Questionnement 1 : À partir d'une étude sur les degrés de similarité entre les données de référence et celles produites dans le cadre moins contrôlé d'Apidae, nous comparerons différentes méthodes computationnelles pour la reconnaissance de similarité. Cette comparaison sera mise en regard de la qualité de différentes qualités de texte (à partir d'une typologie des formes produites par les différents auteurs de ces textes). En d'autres termes, pour reconnaître les objets culturels y a-t-il des approches computationnelles plus adaptées à certaines formes de réécritures de l'objet ?

Questionnement 2 : À partir des données fortement structurées du corpus de référence nous chercherons à identifier les entités nommées et leurs relations au sein du corpus non structuré (Apidae). La perspective de ce traitement automatique des textes est d'en révéler une structuration répondant au paradigme des bases de données relationnelles.

3. Méthodologie

Le dispositif OduS est constitué d'un ensemble logiciel et d'une méthodologie qui permet à des acteurs de produire les données de leur domaine (à des fins de communication et de marketing sur leurs activités). Un consortium est mis en place avec lequel nous définissons les formats de la donnée (structure des bases de données, formulaires de saisie) selon des principes et des concepts issus des sciences de l'information et de la communication. Les formulaires de saisie sont mis à disposition des acteurs du domaine qui les utilisent pour produire et diffuser les informations de leurs activités. Ces formulaires qui s'appuient sur des innovations techno-sémiotiques [3] permettent d'obtenir des données de qualités homogènes et la gouvernance des données qui encadre le dispositif (composé de chercheurs et d'acteurs du domaine) assure le suivi de cette qualité. Ainsi, OduS permet de produire des données structurées et homogène à des fins de communication et de marketing digital pour les acteurs du domaine et à des fins de recherche en informatique et en SHS pour les chercheurs[3].

[2] Le premier domaine d'exploitation est celui du spectacle vivant. Par la suite, la méthodologie pourra être utilisée sur d'autres domaines d'activité socio-économiques afin de constituer de nouveaux corpus mis en correspondances.

[3] Les données produites par OduS sont mises à disposition de tout dispositif tiers en open data. Si ce volet-ci n'est pas l'objet du projet OduS2Apidae, nous soulignons que l'originalité d'OduS est de partager une même base de

En parallèle, les offices du tourisme produisent les données du même domaine depuis la plateforme de mutualisation de l'information touristique *Apidae*. Ces données sont produites dans des formulaires peu adaptés au domaine et par différents utilisateurs aux pratiques de saisie fortement différenciés en fonction des profils de chaque utilisateur et des enjeux de chaque office du tourisme (les raisons de cette pluralité des pratiques de saisie est traité par les collègues du projet DataTour[4]). Les données qui en résultent sont ainsi peu structurées et peu homogènes.
OduS2Apidae est un nouveau module de recherche adossé au dispositif OduS et connecté à la plateforme Apidae dont il récupère les données chaque nuit. Sa fonction est de créer des correspondances (de manière semi-automatisée) entre les données des deux corpus.
La constitution du corpus de recherche à partir d'OduS2Apidae nécessite un travail d'ingénierie pour la réalisation d'un logiciel permettant de créer les correspondances entre les instances d'Apidae et d'OduS (voir les missions détaillées des missions de l'ingénieur d'études en annexe). Ce module comporte une partie d'automatisation des correspondances pour les correspondances non ambigües (sur des données factuelles telles que les coordonnées GPS, la date, l'organisateur de l'événement) et de supervision humaine des correspondances (via des formulaires de présentation des instances).

4. Résultats attendus et caractère innovant de la recherche

Du 1er octobre 2022 au 31 mars 2023, les données culturelles des Hautes-Alpes vont être saisies dans OduS et, dans le même temps, ce sont les offices du tourisme des Hautes-Alpes qui vont saisir les données relatives aux mêmes objets culturels sur la plateforme Apidae[5]. Les saisies des données dans OduS et leur supervision sera réalisée dans le cadre d'un projet financé par le ministère de la culture et le programme européen *Leader*.
Ainsi, le corpus couvrira sur une période de six mois l'ensemble de l'offre culturelle dans le domaine du spectacle vivant sur les Hautes-Alpes. Ceci devrait représenter entre 1500 et 2000 instances pour chacune des deux bases de données (OduS et Apidae). Par la suite, le corpus pourra être élargit au Vaucluse.
Ce corpus constitués de deux bases de données aux formats et contenus différents constituent un matériau riche à la fois pour des recherches en informatique et en SHS.

5. Dimension interdisciplinaire

OduS2Apidae repose sur des données structurées et produites dans un environnement scientifique conçus à partir des Sciences de l'Information et de la Communication. Dans le cadre de nos recherches, nous explorerons deux pistes de recherche interdisciplinaire, mêlant informatique et sciences de l'information et de la communication.

1) étude des similarités textuelles en fonction des formes communicationnelles de l'objet spectacle

La similarité textuelle est un problème encore ouvert en informatique et en linguistique computationnelle. En effet, il n'existe pas une similarité dans l'absolu, mais plutôt des approximations qui peuvent être exprimées en plusieurs niveaux d'abstraction, ayant des particularités bien différenciées. La similarité sémantique entre deux objets textuels (paragraphes, morceaux de texte, phrases voire des morceaux de phrases, etc.) est bien sur l'un des objectifs ultimes de la détection et mise en évidence de la similarité, mais elle est difficilement atteignable. On peut cependant l'approcher via une similarité

données entre chercheurs et acteurs d'un domaine (le spectacle vivant dans le cas d'espèce). Pour les premiers, le dispositif leur assure une donnée de grande qualité pour le développement du domaine. Pour les seconds, le dispositif leur assure une donnée « fraîche », mise à jour quotidiennement et produite dans un cadre scientifique contrôlé.

[4] Gaël Depoorter (LBNC), Yannick Hascoet (Espace-Dev), Lise Renaud (CNE), Eloi Flesch (CNE), Allison Guiraud (CNE), Ouassim Hamzaoui (LBNC), Christina Koumpli (LBNC), Jessica Sainty (LBNC)

[5] Apidae est utilisé par la quasi-totalité des offices des Hautes-Alpes et plus largement par les offices du tourisme de la régions Sud, d'Auvergne Rhône-Alpes et d'Île-de-France.

lexicale par exemple, où les objets textuels sont mis en correspondance non pas au moyen des concepts exprimés, mais par des proximités lexicales entre les mots ou entre groupes des mots. Un moyen intéressant dans le cadre de l'expérimentation à partir des données d'OduS2Apidae sera la mise en correspondance des objets textuels via des méthodes issues de la Physique statistique comme l'énergie textuelle [4,5], qui ne nécessite pas des outils linguistiques dépendants de la langue ni que les phrases soient grammaticalement correctes.

La proximité lexicale peut également être enrichie via une représentation dense des termes utilisés. Ceci peut être fait typiquement avec des modèles neuronaux du type apprentissage profond (deep learning). Ainsi, dans cette représentation chaque terme est plongé dans un ensemble de termes (embeddings) qui lui sont "proches" (lexicalement) mais qui, par leur nombre, leur position et leurs caractéristiques vectorielles calculées dans des vastes corpus, peuvent approcher sémantiquement un contexte pour chaque terme des documents étudiés [6].

Ainsi, nous explorerons les deux genres de proximité, basées sur des méthodes statistiques et neuronales et leur combinaison.

Du point-de-vue des SIC, ce qui nous intéresse ici ce sont les raisons des similarités imparfaites. Qu'est-ce qui fait que les textes sont considérés par ces différentes approches computationnelles comme proches ou éloignés ? Chaque auteur des différents textes d'Apidae se réapproprie les objets culturels à travers différentes modalités de collecte de l'information et de réécriture des contenus. Ces différentes manières de faire donnent lieu à différentes formes de textes pour communiquer les événements culturels. En collaboration avec le projet DataTour2, nous réaliserons une typologie de ces différentes formes communicationnelles et nous verrons en quoi celles-ci déroutent ou non chacune des approches computationnelles étudiées. Cette approche qualitative de la similarité par les SIC devrait permettre de considérer la performance de la linguistique computationnelle en prenant en compte les formes d'écriture de l'objet culturel.

2) étude des entités nommées : de la raison graphique à la raison computationnelle

L'extraction automatique d'entités nommées dans un texte écrit est une des étapes indispensables pour une grande variété de logiciels ayant pour objectif un traitement des données sémantiques exprimées originellement en langage naturel. L'entité nommée est un concept de linguistique computationnelle : il s'agit d'une séquence ininterrompue de mots désignant un lieu (toponyme), une personne ou un personnage (anthroponyme), une structure sociale (ergonyme), etc. Diverses approches et algorithmes ont été développés dans le passé, en s'appuyant sur des règles issues d'une expertise linguistique, ou bien des modèles prédictifs, ou encore sur une combinaison des deux.

Malgré plusieurs défauts, dont leur fameuse "opacité", les modèles prédictifs fondés sur les réseaux de neurones artificiels, appelés "transformateurs", sont aujourd'hui dominants car, pré-entrainés sur de gigantesques quantités de textes, ils permettent d'obtenir une très bonne couverture lexicale et une robustesse satisfaisante. Ils doivent néanmoins être "peaufinés", c'est-à-dire réappris sur des corpus annotés, afin de devenir opérationnels sur des tâches spécifiques comme l'extraction d'entités nommées.

Les données textuelles d'Apidae sont particulièrement intéressantes pour mettre à l'épreuve les approches de l'état de l'art, car ces données sont hétérogènes, peu structurées et peu normalisées. Les données d'OduS fortement structurées constituent un corpus de référence inédit pour procéder aux phases de "peaufinage" demandées par les "transformateurs". Dans le cadre de ce projet, nous mettrons à l'épreuve nos propres approches [7, 8, 9], basées sur la définition de grammaires locales hors contexte. Nous pensons qu'elles permettront de compenser les erreurs commises par les approches considérées aujourd'hui comme l'état de l'art de la reconnaissance automatique d'entités nommées. En outre, au-delà de leur reconnaissance, nous chercherons à identifier automatiquement le type d'associations sémantiques (méronymie, hyponymie, antonymie, etc.) dont elles font l'objet. À partir de ces reconnaissances nous visons à construire automatiquement un graphe de connaissance formel (basé sur le modèle des données d'OduS) permettant de réaliser sur des données faiblement structurées (Apidae), toutes les opérations que l'on réalise habituellement sur des bases de données relationnelles.

La reconnaissance automatisée des entités nommées et leur modélisation à travers un graph de connaissance nous permettront d'analyser la structure informelle des données produites par les différents auteurs d'Apidae. Cette analyse répond à des problématiques de recherche SHS qui s'intéressent à la qualité des données (à l'instar du projet DataTour avec lequel les collaborations sont certaines). Par ailleurs, les textes produits par les offices du tourisme répondent au paradigme informationnel de « la page » (mise en forme de la donnée dans les deux dimensions de l'écran), c'est-à-dire de la raison graphique et, le graph de connaissance répond à celui de la raison computationnelle (la donnée est dissociée de ses supports d'exposition). Ainsi, cette transmutation d'une rationalité à une autre nous permet d'interroger les différentes formes de l'objet culturel en tant que technologies de l'intellect. Il y a peut-être ici, un nouvel espace pour repenser les interactions homme-machine, les traitements computationnels des données et leurs inscriptions à l'écran.

6. Partenariats extérieurs envisagés

Dans le cadre du projet plus large d'OduS, un partenariat a été esquissé avec Apidae. Les résultats de ce travail de recherche permettraient à la plateforme d'une part, d'enrichir ses données (volet informatique) et d'autre part, d'analyser la qualité des données produites par les différents auteurs d'Apidae (volet SIC). En ce sens, nous pensons utile de développer ce premier partenariat et nous pensons que ce renforcement partenarial devrait se faire de concert avec le projet DataTour sous l'égide de la FR Agorantic.

7. Valorisation (si prévue) :

Au-delà d'une utilisation à des fins scientifiques, les résultats de recherche pourront faire l'objet de valorisation auprès de la plateforme Apidae. En outre, ces recherches et les méthodologies qui les sous-tendent pourront être appliquées à d'autres domaines d'activité que le spectacle vivant. Du point-de-vue de leurs applications nos résultats de recherche pourront amener à concevoir des outils permettant :

- D'enrichir des jeux de données en annotant et qualifiant des données culturelles non structurées (dans la perspective d'un open data plus qualitatif)
- Pour agréger des données de différentes sources sans risquer d'afficher des données en doublon (pour un open data plus exhaustif) ; cette possibilité répond déjà à un besoin identifié du projet OduS sur son volet socio-économique.
- Pour analyser le niveau de structuration implicite d'un jeu de données[6] (pour des données textuelles mieux structurées dans une perspective d'open data)

Pour le projet OduS dans son ensemble, nous avons reçu le soutien de l'InSHS qui nous a accompagnés dans les perspectives de valorisation du projet et un premier dépôt de propriété intellectuelle est à l'étude à la Satt Sud-Est.

Budget (€)*		
	Brève description	Montant
Ingénieur d'études (3 mois, 50%)	• Ingénierie : concept° & réalisat° d'OduS2Apidae • Recherche : travaux exploratoires	4.200
Stagiaire (3 mois, 50%)		1.800
Budget demandé à Agorantic		6.000

[6] Quel est le niveau de redondance, entre les différentes instances du corpus, de la structuration implicite d'un champ textuel. Par exemple, la structure de présentation d'un spectacle respecte-t-elle toujours la même structure : genre, chapeau, texte descriptif, distribution, etc. ou la structuration implicite change-t-elle à chaque instance ?

Bibliographie

[1] Goody, J. (1979). *La raison graphique. La domestication de la pensée sauvage*. Paris : Les Editions de Minuit.

[2] Bachimont, B. (2004). *Arts et sciences du numérique : Ingénierie des connaissances et critique de la raison computationnelle* (Habilitation à diriger des recherches). Université de Compiègne, Compiègne.

[3] Flesch, E. (2019). OduS, dispositif et corpus. Dans I. Roxin, F. Tajariol, I. Hosu et N. Pélissier (dir.), *Information, Communication et Humanités Numériques. Enjeux et défis pour un enrichissement épistémologique* (p. 157 - 175). Cluj Napoca : accent.

[4] S Fernandez, E SanJuan, JM Torres-Moreno, Textual energy of associative memories: Performant applications of enertex algorithm in text summarization and topic segmentation, MICAI, 861-871

[5] J.M Torres-Moreno, Automatic Text Summarization. Wiley, London 2014.

[6]. Predicting the Semantic Textual Similarity with Siamese CNN and LSTM, E Linhares Pontes, S Huet, AC Linhares, JM Torres-Moreno, arXiv:1810.10641 1810 (10641)

[7] Murata J., Carrette R. & Jourlin P. (2021). SIDRES: A Novel Annotation Tool For The Automatic Detection of Semantic Entities. Actes de Traitement Automatique des Langues Naturelles, Lille, France. pp.15-17. ⟨hal-03265913⟩

[8] JOURLIN P. (2022) Disambiguation for the Classification of Lexical Items. France, Patent n° : EP3937059A1. ⟨hal-03598242⟩

[9] Jourlin P. (2022). SIMI : un système de suggestion de littérature médicale. Actes de Traitement Automatique des Langues Naturelles, Avignon, France.

Annexe - Missions de l'ingénieur d'études

L'ingénieur d'études réalisera les éléments logiciels du dispositif de mise en correspondance des données OduS et Apidae. Le volume horaire est estimé à un peu plus de 200h, soit 3 mois à mi-temps. Pour ce poste, Eloi Flesch qui a conçu et réalisé OduS est présupposé à la réalisation de ce travail d'ingénierie. Cette ingénierie pouvant être menée en parallèle de celle du projet DataTour2, ce mi-temps viendra donc en complément du mi-temps qui est prévu sur le projet DataTour2.

Préparation des données d'OduS

- Création de la structure de la donnée pour le spectacle vivant sur le Briançonnais selon la méthode développée dans le cadre du projet OduS
- Configuration d'OduS (création des formulaires et structuration de la base de données)

Réalisation de tables spécifiques pour la gestion des correspondances

- Table de scoring entre les objets d'Apidae et d'OduS
- Table de mise en correspondance
- Table de configuration des paramètres

Mise en correspondance automatique entre les objets d'Apidae et d'OduS

- Réalisation d'un scoring quotidien entre les objets Apidae et OduS à partir de données factuelles (dates, heure, coordonnées GPS, contacts téléphoniques, mails, etc.) Ces données devront être préparées pour être au même format (ex : Saint-Paul -> St Paul, 06.00.00.00.00 -> 06 00 00 00 00, etc.)
- Création automatique de la correspondance si le scoring est supérieur à un certain seuil configurable

Traitement manuel des correspondances en souffrance

- Création d'une interface de présentation des objets en souffrance et de mise en correspondance (sélection > validation)
- Création d'une interface de présentation des correspondances (avec filtres) à des fins d'éventuelle modification

Création d'une interface de configuration

- Configuration des pondérations de scoring (afin d'être adaptés au fil du projet)
- Définition des seuils de scoring pour les correspondances