APPEL À PROJETS 2022 – DONNÉES & OPEN DATA

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

Titre	DataTour 2
Acronyme	DataTour 2
Nom du/des porteur(s)	Gaël Depoorter, Yannick Hascoët
Laboratoires associés	LBNC, Espace
Budget demandé	8.450
Résumé *Max. 1 000 caractères espaces compris*	Dans un contexte d'ouverture des données à des fins de diversification de leurs usages, la qualité des données est au cœur des enjeux d'exploitabilité. En 2017, l'État lance DataTourisme, une plateforme où convergent les données touristiques des territoires pour y être agrégées puis mises à disposition pour favoriser l'innovation et de nouveaux usages numériques dans le secteur touristique. Pour autant 5 ans après, l'hétérogénéité de la qualité de ces données rend difficile leur exploitabilité. Quelles sont les causes de cette qualité hétérogène ? C'est en descendant au plus près des producteurs de ces données, en analysant les circonstances dans lesquelles sont prises les pratiques de saisie, et en suivant le parcours de remontée des données que nous cherchons à décrire les entraves à la promesse d'un open data efficient et au service du développement touristique. Ainsi, ce projet s'inscrit dans le troisième objectif de cet AAP.

1. Contexte, positionnement, objectif(s)

Contexte

D'abord destinée aux données des administrations publiques, l'injonction à l'open-data s'étend à d'autres secteurs et notamment aux données touristiques ; ces dernières sont désormais soumises à des obligations légales d'ouvertures[1].

Dans ce cadre, l'État et la fédération nationale *Tourisme et Territoires* ont développé la plateforme DataTourisme afin d'agréger les données produites par les différents systèmes d'information touristiques (SIT) des différents territoires nationaux[2]. L'une des missions de la plateforme est de standardiser les données issues des 33 bases de données qui se répartissent le territoire. Ceci étant, la convergence des données issues des différents SIT vers la plateforme de DataTourisme ne se fait pas sans difficultés. Si la remontée est bien effective c'est l'hétérogénéité de leurs qualités qui pose problème. Cette hétérogénéité constitue une entrave à l'innovation dans le champ de l'open data (le projet même de DataTourisme) et la promesse d'un open data touristique efficient s'en trouve contrarié.

Positionnement

Le contexte présenté succinctement permet de saisir pourquoi la standardisation des contenus produis par les différents SIT est un objectif majeur et affiché de DataTourisme[3]. Pour autant, nos premiers entretiens montrent que, soit les organisations ne font que peu de cas de l'enjeu de l'open data, soit elles mettent en place des stratégies de résistance à l'injonction de l'open data afin de garder une partie

[1] https://www.entreprises.gouv.fr/files/files/directions_services/tourisme/datas/Guide_juridique_open_data_vDEF07062018.pdf
[2] La plateforme vient d'être transféré à ADN Tourisme.
[3] (https://info.datatourisme.fr/fonctionnement/ontologie/

de la valeur de la donnée. Quand certains cherchent à concevoir une ontologie qui mettrait en ordre de bataille l'ensemble des acteurs du tourisme qui agissent sur la donnée (Soualah-Alila, Coustaty, Faucher, et Wannous, 2016), notre positionnement consiste à interroger les raisons du difficile avènement de l'open data dans le domaine touristique. Dans le prolongement des travaux d'Éric Dagiral et Sylvain Parasis, qui rappellent que « les données sont toujours des constructions sociales et politiques » (Dagiral et Parasie, 2017), notre approche n'est pas techniciste mais s'intéresse à la donnée en tant que production sociale et anthropologique, politique, économique et normative, technique et sémiotique.

Objectif général :

Lors de la première phase du projet DataTour (AAP Agorantic 2021) nous nous sommes intéressés à l'un des deux plus importants SIT : la plateforme Apidae. Nous avons mené des entretiens exploratoires et réalisé un premier travail d'analyse des « fiches » d'objets touristiques (au format Json) produites depuis la plateforme Apidae. Cette première phase de la recherche nous a permis de faire le constat de deux entraves majeures à l'homogénéisation des données au niveau de DataTourisme.

- La première entrave est celle de la convergence des formats qui se joue au niveau des configurations techniques afin de « préparer » les données au format unique de DataTourisme (mapping). Pour le dispositif Apidae, la décision de faire remonter telle donnée dans tel format revient aux acteurs du tourisme et se joue à un échelon local ou départemental. Ceci engendre une pluralité de configuration rendant difficile la mise en ordre de bataille des acteurs du tourisme pour une donnée homogène et unifiée. Ainsi, ces configurations sont en proie à des enjeux techniques (possibilités offertes par les SIT), juridiques, économiques (concurrentiels), politiques, etc.
- La seconde entrave se joue à l'endroit des pratiques de saisie. Ces pratiques sont « configurées » par différents enjeux autour de la donnée : socio-politiques, sociotechniques, socio-professionnels, techno-sémiotiques, etc.

Ainsi, notre objectif général est de mettre au jour les raisons sociales, politiques, juridiques, techno-sémiotiques et sociotechniques qui participent à nourrir et à légitimer les pratiques qui aboutissent à ces deux entraves.

Objectif 2022 :

Pour cette seconde phase (2022), nous nous concentrons sur la deuxième entrave en voulant appréhender par l'enquête les multiples raisons qui engendrent des données de qualités hétérogènes. Nous considérons cette hétérogénéité comme découlant de la diversité des praticiens qui utilisent le SIT Apidae et de la pluralité des contextes professionnels dans lesquels ces pratiques s'inscrivent (différents offices du tourisme et différents enjeux qui leurs sont liés.

La question de l'ouverture des données et de leur apprêtement à des fins d'open data a été documentée dans le cadre des données publiques administratives (Goeta, 2016). Mais à notre connaissance, aucune recherche n'a porté sur la qualité des données touristiques depuis leurs causes. Dans le cadre des pratiques de l'open data relatives aux données des administrations publiques, une panoplie d'opérations est mise en œuvre pour rendre exploitables les données dans d'autres contextes desquelles elles sont issues.

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

Apidae est un CMS (Contents Management System) conçu pour le secteur touristique. Il offre ainsi un « même moule » d' « écriture sous contrainte » (Jeanne-Perrier, 2005) à l'ensemble de ses utilisateurs. Il serait donc tentant de penser qu'un formulaire ainsi unifié permettrait de produire des données aux qualités homogènes. Toutefois, l'étude exploratoire que nous avons menée avec trois étudiantes du master Gouvernance des données montre une hétérogénéité des contenus produits au sein d'Apidae

par de multiples praticiens de la donnée touristique. C'est ce que souligne Patrice Flichy lorsqu'il écrit que « l'hétérogénéité des acteurs et des pratiques résiste en permanence au projet d'homogénéisation des systèmes d'information. [...] À partir de la même vue des données, on peut observer des pratiques fort différentes. » (Flichy, 2013, p. 81) Cette « résistance » a fait l'objet de travaux en sociologie des usages (Jouët, 2000) et rappelle aussi à la créativité culturelle et la liberté manifestées par des acteurs ordinaires, au quotidien, qu'observe Michel De Certeau (1980) Dans cette perspective et en s'intéressant aux « Technologies de l'Information et de la Communication » au sein des organisations, Durampart (2007) écrit que « les TIC sont bien situées à l'intersection de l'individuel et du collectif, de la technique et du social, de la norme, de la rationalisation et d'une appropriation plus réactive et inventive ».
Dès lors, comment ces différentes tensions agissent sur les praticiens de la donnée du dispositif Apidae ? Comment les qualités des données qu'ils produisent en dépendent ? Et comment ces qualités hétérogènes mettent en échec le commun de la donnée et l'avènement d'un open-data efficient ?
Ce questionnement s'inscrit dans le troisième des quatre objectifs de cet AAP qui s'intéresse à « la qualité des données (fiabilité, homogénéité) en tant que condition d'un OpenData efficient », à « ses modalités de recueil », à « l'obstruction des sources » et aux « modes de gouvernance pour la production et la diffusion » de ces données.

3. Méthodologie

Notre méthodologie est multifocale, pluridisciplinaire, et fait appel à différentes techniques d'enquête.
Notre méthodologie est multifocale dans le sens où d'une part, nous nous intéressons à la donnée depuis différents point-de-vues : 1) celui des acteurs de son économie (là où sont décidés ses usages, ses modalités sociotechniques de collecte, de production et de circulation) ; 2) celui du producteur de la donnée (là où se réalise, en pratique, dans un contexte professionnel quotidien l'activité même de l'écriture de la donnée) ; 3) celui de la donnée et de ses propriétés intrinsèques (leurs qualités résultant des pratiques situées). Et d'autre part, dans le sens où nous l'observons depuis différentes « profondeurs de champs » : de l'office du tourisme à DataTourisme.
En outre, partant d'un état de l'art quasi inexistant sur les causes multifactorielles de l'hétérogénéité des données, nous chercherons à faire émerger des hypothèses depuis le terrain, en nous appuyant sur les principes de la théorisation ancrée (Paillé, 2011). Nos différentes techniques d'enquête pour appréhender les données (analyse des bases de données, entretiens, observations ethnographiques, etc.) doivent nous permettre de définir les contextes dans lesquels les pratiques s'inscrivent, de caractériser ces pratiques et d'analyser les qualités qui en résultent.

4. Résultats attendus et caractère innovant de la recherche

Résultats attendus :
Nous cherchons à montrer comment les spécificités d'un office du tourisme (sociologiques, managériales, institutionnelles, sociopolitiques, économiques, démographiques, organisation techno-sémiotiques et sociotechnique de la donnée, etc.) ainsi que les spécificités des pratiquants de la donnée avec Apidae (pratiques professionnelles, littératie numérique, représentations et imaginaires) prédéterminent en partie la qualité des données qui seront produites au sein de tel office du tourisme ou par tel praticien. Nous visons à pouvoir définir des causes de l'hétérogénéité des qualités d'une donnée en fonction de leurs contextes (individuels et collectifs) de production.
Caractère innovant de la recherche :
S'il est acquis que la qualité des données constitue un enjeu majeur pour l'open data et le développement d'applications à partir de ces données, ce qui n'est pas résolue ce sont les causes

anthroposociales qui mènent à des données difficilement exploitables informatiquement. L'enjeu de la qualité des données est souvent abordé depuis la pertinence de l'ontologie qui représenterait le domaine ou depuis son nettoyage a posteriori mais peu de travaux s'intéressent à la donnée prise dans des enjeux sociaux. C'est en prenant au sérieux le fait que les données ne sont pas données et qu'elles sont belles et bien prises dans des pratiques et des représentations sociales, dans des enjeux politiques, économiques et sociaux que nous pouvons aborder la qualité de la donnée non plus d'un point-de vue techniciste mais d'un point-de-vue social et culturel.
À notre sens, il y a là un questionnement épistémologique à la croisée de l'informatique et des SHS qu'il serait fertile de questionner au sein de la FR Agorantic : comment peut-on *penser* les données à l'intersection de leur valeur anthroposociale et de leur calculabilité.

5. Dimension interdisciplinaire

En prenant en compte les différents éléments de contexte (à l'échelle individuelle et collective), le projet DataTour est nécessairement interdisciplinaire. En 2022 nous avons resserré l'équipe pour la réalisation des enquêtes à partir de nos travaux exploratoires de 2021. En 2023, les données d'enquête (entretiens, observations, données Apidae) seront traitées par les différentes disciplines et chercheurs de DataTour : Gaël Depoorter (sociologie politique, LBNC), Yannick Hascoët (géographie et ethnographie, EspaceDev/CNE), Lise Renaud, Allison Guiraud et Eloi Flesch (sciences de l'information et de la communication, CNE), Christina Koumpli (droit, LBNC).
En outre, l'équipe est amenée à s'étendre notamment avec Jimmy Merlet (économètre, LBNC) qui traitera nos données avec des modèles économétriques afin de vérifier nos hypothèses de corrélation et permettre ainsi de mettre au jour des variables déterminantes sur la manière dont sont saisies les données et la qualité qui en résulte. Par ailleurs, DataTour est désormais en lien avec le projet OduS (Eric Triquet, Juan-Manuel Torres-Moreno, Eloi Flesch) qui nous met à disposition les données d'Apidae et des outils et compétences pour leur analyse.

6. Partenariats extérieurs envisagés

Le lien avec le projet OduS (qui récupères les données d'Apidae chaque nuit) nous permet d'avoir un premier partenariat avec la plateforme Apidae que nous entendons développer au fil de ce projet. En fonction des opportunités, nous étudierons la possibilité de nouveaux partenariats avec ADN Tourisme (DataTourisme), Vaucluse Animation et Datactivist (qui a pour mission le développement de l'open data sur différents secteurs d'activité).

7. Valorisation (si prévue) : déclaration d'invention permettant de valoriser un savoir-faire, une base de données ou un logiciel

Les résultats de la recherche constitueront un socle scientifique pour l'éventuelle conception d'un logiciel d'analyse semi-automatisée de données et pour préconiser des évolutions des pratiques pour une homogénéisation des données entre les différents offices de tourisme. L'enjeu est celui de l'exploitabilité des données touristiques au-delà des dispositifs et nejeux locaux afin de donner de la valeur aux données dans un contexte d'open-data. Ce prolongement pourrait aboutir à une déclaration d'invention destiné à un secteur à forte valeur économique.

Budget (€)		
	Brève description	Montant
Ingénieur d'études (5 mois, 50%)	Entretiens et observations, analyse des dispositifs et de la qualité des données	7.750
Stages (2 stages de 3 mois)	Entretiens et observations	1.800
	Conception d'un outil de data-visualisation	1.800
Prestation	Transcription d'entretiens	4.000
Missions	Déplacement entretiens : Offices du tourisme (Vaucluse), Datatourisme (Paris), Apidae (Lyon)	2.500
Consommables, petits matériels**	Achat de licences (data-visualisation)	400
Budget total		18.250
Co financement	Projet DataTour 1 (AAP Agorantic 2021)	3.000
	Projet DataEthno (AAP CR - Axe identitaire)	6.800
Budget demandé à Agorantic		8.450

Bibliographie

Baudot, P.-Y., Marrel, G. et Nonjon, M. (2015). Encore une révolution informatique ? Open et big data dans les organisations administratives. *Informations sociales*, *191*(5), 8-18.

Dagiral, É. et Parasie, S. (2017). La « science des données » à la conquête des mondes sociaux : Ce que le « Big Data » doit aux épistémologies locales. Dans P.-M. Menger et S. Paye (dirs.), *Big data et traçabilité numérique : Les sciences sociales face à la quantification massive des individus* (p. 85-104). Paris : Collège de France.

De Certeau, M. (1980). *L'invention du quotidien*. *1 Arts de faire*, et *2 Habiter, cuisiner*, Paris : Folio.

Durampart, M. (2007). Les TIC et la communication des organisations : Un dispositif révélateur des émergences ambivalentes de nouvelles formes organisationnelles. *Communication et organisation*, (31), 164-177.

Flichy, P. (2001). *L'imaginaire d'Internet*. Paris : La Découverte.

Flichy, P. (2013). Rendre visible l'information : Une analyse sociotechnique du traitement des données. *Réseaux*, *178-179*(2), 55-89.

Goeta, S. (2016). *Instaurer des données, instaurer des publics : Une enquête sociologique dans les coulisses de l'open data* (Sociologie). Télécom ParisTech, Paris.

Jeanne-Perrier, V. (2005). L'écrit sous contrainte : Les Systèmes de management de contenu (CMS). *Communication & Langages*, *146*(1), 71-81.

Jouët, J. (2000). Retour critique sur la sociologie des usages. *Réseaux*, *18*(100), 487-521.

Le Deuff, O. (2012). Littératies informationnelles, médiatiques et numériques : De la concurrence à la convergence ? *Études de communication*, (38), 131-147.

Paillé, P. (2011). L'analyse par théorisation ancrée. *Cahiers de recherche sociologique*, (23), 147-181.

Soualah-Alila, F., Coustaty, M., Faucher, C. et Wannous, R. (2016). Projet TourinFlux : Apport des Technologies du Web Sémantique pour la Gestion des Données du Tourisme (pp 12). Communication présentée au Colloque AsTRES : Association Tourisme Recherche et Enseignement Supérieur, Université de Bretagne occidentale.

Souchier, E., Jeanneret, Y. et Le Marec, J. (Dirs.). (2003). *Lire, écrire, récrire : Objets, signes et pratiques des médias informatisés*. Paris : Bibliothèque Publique d'Information.