Synthèse des résultats des expérimentations des outils OSC au Tubà et à la Péniche Livrable n°4

Valentyna Dymytrova, Isabelle Hare, Marie-France Peyrelong, Valérie Larroche, Françoise Paquienséguy

elico

Équipe de recherche de Lyon en sciences de l'information et de la communication

Synthèse des résultats des expérimentations des outils OSC au Tubà et à la Péniche

Livrable n°4

Rédacteurs :

Valentyna Dymytrova, Postdoc

Isabelle Hare, MCF

Marie-France Peyrelong, MCF

Contributeurs :

Valérie Larroche, MCF

Françoise Paquienséguy, PR

Juin 2018

INTRODUCTION

Porté par le laboratoire Elico (EA 4147), ce livrable n°4 est associé à la tâche 2.4 « Feedback on developers' and designers' uses of the platform with recommendations for the design of services or applications". Il analyse des situations d'usages des outils développés dans le cadre de l'ANR OpenSensingCity par des participants des expérimentations mises en place. La première expérimentation réalisée par l'équipe du projet a eu lieu le 17 mars 2017 à Lyon, la deuxième s'est déroulée le 13 décembre 2017 à Grenoble.

Ce livrable est organisé en trois parties. La première décrit le contexte des expérimentations et explicite les protocoles méthodologiques déployés. La deuxième partie rend compte des résultats des expérimentations menées au Tubà et à La Péniche en mettant en avant la place des outils OSC dans trois étapes de la chaîne de traitement, à savoir la récupération, la transformation et la réutilisation des données. La troisième partie met en perspective les outils développés par rapport aux pratiques des participants des expérimentations. Enfin, la synthèse finale met en exergue des points à prendre en compte dans la conception des outils facilitant le traitement des données ouvertes.

1. CONTEXTE ET ÉTAT DE L'ART

Issue des sciences de la nature, la méthode expérimentale consiste à « *manipuler une ou plusieurs variables pour mesurer leur impact sur la variable à expliquer, en isolant au maximum les facteurs externes pouvant perturber cette relation* »[1]. Dans cet objectif, elle recrée un environnement pour comprendre un fonctionnement ou trouver une solution à un problème donné. L'expérimentation réunit ainsi « *les conditions nécessaires pour s'assurer de la reproductibilité du fonctionnement du dispositif* »[2.]

Centrale pour les sciences de l'ingénieur, l'expérimentation rassemble les individus, un contexte social et des outils à tester. Elle relève d'une recherche tâtonnante qui peut

1 Livian, Y. (2015). Initiation à la méthodologie de recherche en SHS. <halshs-01102083>.
2 Gentès, A. & Jutant, C. (2011). Expérimentation technique et création : l'implication des utilisateurs dans l'invention des médias. Communication & langages, 168,(2), 97-111. doi:10.4074/S0336150011012087.

déboucher sur un échec, toujours significatif pour la réflexion[3]. La validité de l'expérimentation est toutefois garantie par le protocole et sa bonne mise en place. En ce sens, elle peut être réitérée ou refaite en modifiant certains critères[4].

Dans le cadre de notre ANR, les deux expérimentations interviennent comme des moments inédits, qui ont permis de tester des hypothèses et ont supporté le processus de conception en confrontant les outils en cours de développement à des réelles situations d'usage. La préparation des expérimentations a aussi stimulé l'avancement du développement des outils et la préparation des jeux de données par des chercheurs en informatique et les partenaires industriels. Outre cela, les expérimentations ont assuré une véritable dynamique interdisciplinaire et multipartenaire entre les membres du projet. L'élaboration commune du protocole de l'expérimentation y a beaucoup contribué.

2. MÉTHODOLOGIE

2.1. PRÉSENTATION GÉNÉRALE

Dans notre cas, les expérimentations OSC sont intervenues en cours du processus de conception des outils et non pas à la fin, à titre d'évaluation des produits conçus. Nous sommes ainsi face à un cycle de la démarche expérimentale car les résultats tirés de la première expérimentation ont permis de mettre en place la deuxième.

Chaque expérimentation a articulé trois temporalités :

- celle d' « avant » avec la mise en place de l'expérimentation qui consistait dans la préparation des données et des outils, la formulation des hypothèses et leur opérationnalisation sous forme d'élaboration du protocole ;
- celle de « pendant » l'expérimentation qui renvoyait au moment de la passation de l'expérimentation et au travail d'observation et de collecte de matériaux ;

3 Dumouchel, S. (2015). « Comment et pourquoi construire un protocole en SHS ? », Digital Humanities à l'IHA, https://dhiha.hypotheses.org/1519.
4 Milhabet, I., Thérouanne, P. (2016). « L'approche expérimentale en SHS », présentation au séminaire de l'axe 2 de la MSH Sud-Est : http://mshs.unice.fr/wp-content/uploads/2016/07/MSHS-Axe-2-Approche-experimentale-Milhabet-Therouanne.pdf.

- celle d' « après » l'expérimentation qui consistait dans l'analyse des matériaux collectés, la formulation des résultats et l'évaluation du dispositif d'expérimentation.

2.2. PROTOCOLE D'OPENSENSINGCITY CHALLENGE, 17 MARS 2017, TUBA

La préparation de chaque expérimentation a fait l'objet de plusieurs réunions d'équipe et conférences téléphoniques. Les discussions ont porté plus précisément sur le protocole d'expérimentation. Nous avons collectivement défini les cibles, les étapes de chaque expérimentation, les rôles dévolus à chacun et les résultats attendus à chaque étape. Les protocoles de chaque expérimentation ont été formalisés sous forme d'un document partagé avec l'ensemble de l'équipe.

Partenariat avec Tubà, living lab de la métropole de Lyon

La première expérimentation, OpenSensingCity Challenge s'est tenue le 17 mars 2017, de 9h à 17h dans les locaux de Tubà, le living lab de la métropole de Lyon[5]. Situé dans le quartier d'affaires Lyon Part-Dieu, Tubà est porté par l'association Lyon Urban Data qui regroupe une quarantaine de partenaires publics et privés parmi lesquels des collectivités, grands groupes, PME, startups, laboratoires de recherches et citoyens. Inauguré en novembre 2014, Tubà cherche à favoriser l'innovation, l'incubation et le développement de services urbains dans une démarche collaborative et participative s'appuyant sur les données numériques privées et publiques. Pour cela, il réunit dans un espace de 600 m^2 une équipe composée d'une directrice et de quatre chargés de mission et plusieurs porteurs de projet relatifs à la smart city. Les locaux de Tubà proposent aussi un espace de coworking gratuit et ouvert à tous.

Objectifs de l'expérimentation

L'expérimentation OpenSensingCity Challenge[6] visait une confrontation stimulante des résultats et des questionnements entre différentes catégories d'acteurs (chercheurs et professionnels des data) afin de dégager des réalisations, des pistes : 1/ quant à la pertinence de scénarios d'usages de réutilisation des données ouvertes par des

5 http://www.tuba-lyon.com.

6 L'équipe d'organisation de l'expérimentation OpenSensingCity Challenge comprenait : Noorani Bakerally (École des mines de Saint-Étienne), Valentyna Dymytrova (Science Po Lyon), Amélie Gyrad (École des mines de Saint-Étienne), Valérie Larroche (IUT Lyon 3), Maxime Lefrançois (École des mines de Saint-Étienne), Marie-Francec Peyrelong (ENSSIB) et Antoine Zimmermann (École des mines de Saint-Étienne).

professionnels de la donnée ou des applications 2/ quant aux problèmes rencontrés, ou résolus que ceux-ci rencontrent en situation de réutilisation de données ouvertes 3/ quant aux outils OSC développés par l'EMSE. Une page web spécifique a été créée en vue de la préparation de l'expérimentation : http://opensensingcity.emse.fr/tuba/.

Participants

À partir des résultats de la première enquête de terrain menée par Elico (cf. Livrable 1[7]), deux publics spécifiques ont été identifiés pour l'expérimentation à Tubà : développeurs et data scientists. Les participants ont été recrutés via les personnes précédemment interviewées dans le cadre de deux enquêtes de terrain, le carnet d'adresse de Tubà et la plateforme Meetup où une annonce spécifique a été créée.

Au total, 11 professionnels des données ont participé à l'expérimentation. Ce sont des développeurs, des data analystes et des data scientists (population masculine de 24 à 40 ans) qui travaillent pour des entreprises spécialisées en traitement et analyse des données (Datalyo[8], Dascils, Keyrus[9]), l'agence de design de services Atrioom10, une entreprise suédoise de conception de plateformes de vente Universal Avenue, une entreprise américaine de conception de solution et d'application de services Wingz[11], un data scientist de la Métropole de Lyon et deux doctorants en informatique de l'INSA de Lyon.

Les participants ont travaillé en groupe de 2, 3 ou 4 personnes.

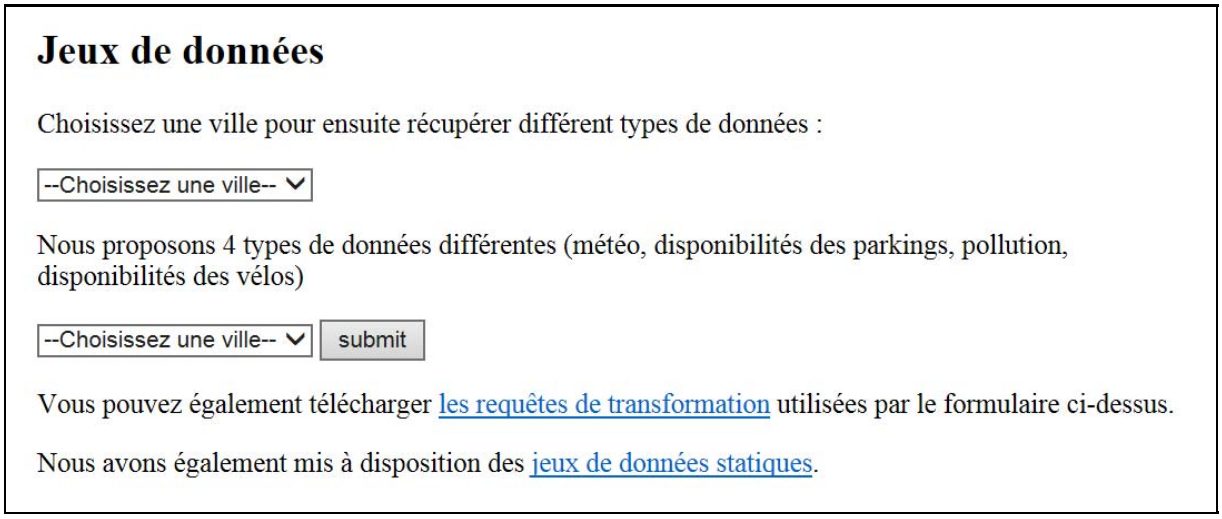
Données et outils mis à disposition

L'hypothèse principale de la journée consistait à vérifier si l'usage d'un modèle commun de présentation de données ouvertes (RDF) et la possibilité d'un unique point d'accès aux données (portail Antidot OSC) faciliteraient l'exploitation des données ouvertes analogues issues de différents territoires.

7 Paquienséguy, F., Larroche, V., Peyrelong, M-F., Vila-Raimondi, M., Dymytrova, V. (2016). Synthèse des résultats de l'enquête auprès de ré-utilisateurs de données ouvertes. Livrable 1. Elico. <hal-01432124>.
8 http://www.datalyo.com/.
9 http://www.keyrus.fr/.
10 http://www.atrioom.fr/.
11 https://www.wingz.me/.

De fait, l'équipe de l'EMSE a mis à disposition des participants les données ouvertes transformées dans un modèle commun (RDF) et le format JSON-LD (JSON for Linking Data) avec l'outil SPARQL-Generate qu'ils ont développé. L'outil lui-même n'a pas fait l'objet de test, seules les requêtes de transformation étaient disponibles à partir de la page dédiée à l'expérimentation. Il s'agissait de quatre types de données différentes (météo, disponibilités des parkings, pollution, disponibilités des vélos) issues des portails des données ouvertes de cinq métropoles françaises : Paris, Lyon, Strasbourg, Nantes et Rennes.

Figure 1. Capture d'écran de la page web de l'expérimentation OSC.



À son tour, l'entreprise Antidote a proposé une plateforme expérimentale d'indexation de données pour la navigation et la recherche qui comprenait un service Web avec une documentation de l'API. Dans l'état de prototype, celle-ci indexait un nombre limité de données[12].

Scénario

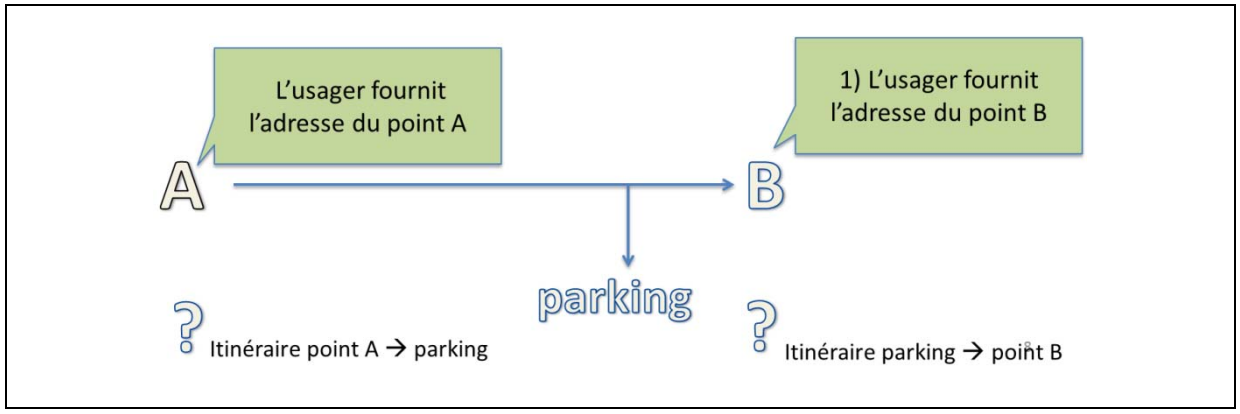
Afin de tester l'hypothèse principale de l'expérimentation, nous avons proposé aux participants d'exploiter des données mises à disposition dans un modèle unique (RDF) et de créer à partir de ces données une application, une visualisation ou tout produit qui

12 http://eval02.partners.antidot.net/default/osc/tuba/results.html#!search.

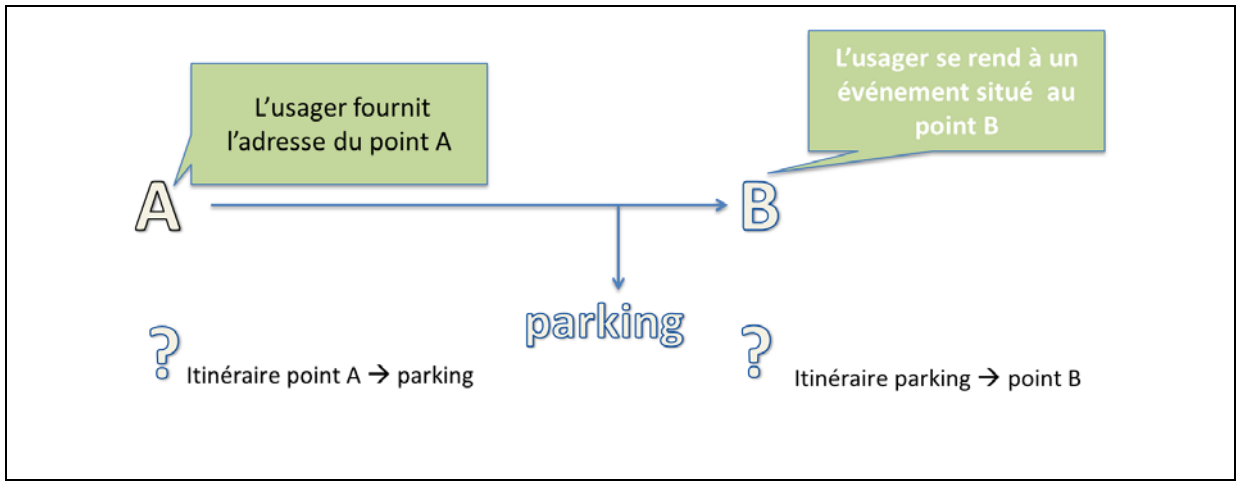
favoriserait l'aide au stationnement en milieu urbain[13]. La création devait être utilisable dans des villes différentes.

Ce scénario a été décliné en trois options d'une complexité croissante :

Option 1 : L'usager se rend d'un endroit A à un endroit B en passant par un parking. Les adresses de deux points sont disponibles.

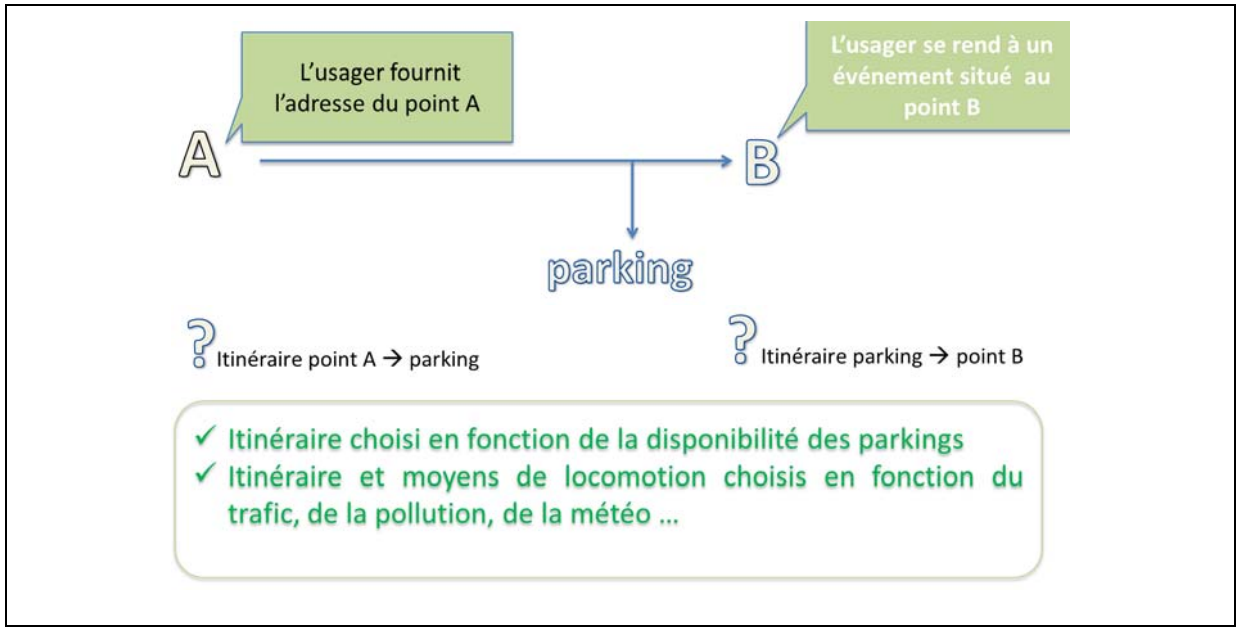


Option 2 : L'usager se rend d'un point A à un événement situé au point B en passant par un parking. Seule l'adresse du point A est disponible.



Option 3 : L'usager se rend d'un point A à un événement situé au point B. Il choisit son itinéraire en fonction de la disponibilité des places de parking et du trafic, de la pollution, de la météo.

13 http://www.emse.fr/~zimmermann/OpenSensingCity/tuba-2017-03-17.html.



Matériaux collectés

Répartis en quatre groupes, les participants ont été invités à restituer les résultats de leur travail en fonction de plusieurs éléments : la réalisation, les données utilisées, la justification du choix du scénario, les étapes essentielles de la réalisation, les outils utilisés, les points de négociation dans le groupe et l'avis sur les données et les outils mis à disposition. Par ailleurs, nous avons demandé à chaque groupe de réaliser des captures d'écran avec le logiciel Camtasia[14].

À l'issue de la journée, nous avons constitué un corpus multimodal et multi source des données :

- Photographies prises lors de la journée (organisation de l'espace du challenge, dispositions des participants) ;
- Enregistrements audio des interactions entre les participants de chaque groupe ;
- Vidéos des présentations orales des réalisations finales ;
- Captures d'écrans des étapes principales d'avancement de chaque groupe ;
- Notes d'observation.

14 https://www.techsmith.com/download/camtasia/.

Dans la semaine suivant l'événement, nous avons effectué les entretiens post expérimentation d'une heure environ avec six participants de façon à avoir au moins un représentant de chaque groupe de travail.

2.3. PROTOCOLE D'ATELIER DÉVELOPPEUR OPENSENSINGCITY, 13 DÉCEMBRE 2017, LA PÉNICHE

La préparation de l'expérimentation a fait l'objet de plusieurs conférences téléphoniques et réunions de l'équipe des organisateurs. Elle a stimulé l'avancement du développement des outils par des chercheurs en informatique et l'indexation des jeux de données par Antidot. L'équipe Elico étant également investie dans l'indexation à travers la mise en place d'une taxonomie des jeux de données liées à la mobilité commune à plusieurs portails métropolitains de l'open data.

Partenariat avec La Péniche, la Coop Infolab à Grenoble

L'expérimentation s'est tenue le 13 décembre 2017 dans les locaux de La Péniche[15], une entreprise coopérative, fondatrice de la Coop Infolab[16]. Celle-ci est construite sur le modèle d'un FabLab, lieu où se côtoient développeurs, usagers, entrepreneurs, chercheurs, associations, grands comptes, créatifs et collectivités locales pour imaginer et designer de nouveaux services numériques bâtis sur les données. Une équipe permanente de quatre chefs de projets et de deux animatrices accueille et sensibilise les publics intéressés par les questions de données. La Péniche héberge également plusieurs résidents qui développent des projets sur des thématiques des données numériques, usages numériques collaboratifs, innovations sociales, transition énergétique et industries créatives.

Objectifs de l'expérimentation

L'objectif de cette deuxième expérimentation était de tester des outils développés dans le cadre du projet en concevant une extension d'une application existante, basée sur l'exploitation des données ouvertes ainsi que des outils du Web sémantique. L'hypothèse principale consistait à vérifier si les outils développés facilitaient l'utilisation de l'open data

15 https://www.la-coop.net/la-scop-la-peniche/
16 http://infolabs.io/sites/default/files/Presentation%20La%20Coop%20Infolab%20Grenoble.pdf.

en particulier lorsqu'une application devait fonctionner avec des données de multiples portails de données ouvertes urbaines.

Participants

Les outils développés s'adressent principalement à des professionnels des data dont les pratiques et les cadres d'usage des données ont fait l'objet d'enquête de terrain spécifique (cf. Livrable 3[17]). Pour l'expérimentation à La Péniche, nous avons cherché en particulier des développeurs et des professionnels familiers des technologies du web sémantique puisque les outils développés par les chercheurs de l'EMSE se basent sur ces technologies.

Les participants ont été recrutés via le carnet d'adresse professionnel des chercheurs de l'EMSE et le réseau des contacts de La Péniche, bien insérée dans le paysage numérique grenoblois.

Au total, sept professionnels de horizons variés ont participé à l'expérimentation : deux ingénieurs R&D de la société Smart Origin[18], spécialisée en édition des logiciels ; un ingénieur de recherche Orange Labs, un chargé de mission « Données ouvertes » de la Métropole de Grenoble, un développeur de l'université Grenoble Alpes (UMS3758 Recherche-Infrastructure de CAlcul Intensif et de Données (GRICAD) et deux étudiants en informatique de l'université Grenoble Alpes. Les motivations à participer à la journée étaient différentes : tous y voyaient un intérêt professionnel que ce soit pour mieux comprendre les cadres d'application du web sémantique ou pour se familiariser avec les usages et la manipulation de l'open data.

Les participants ont travaillé en groupe de 2, 3 ou 4 personnes.

Données et outils mis à disposition

Pour cette deuxième expérimentation, plusieurs données et outils ont été préparés et mis à disposition. L'ensemble est listé sur la page web dédiée[19].

17 Dymytrova, V., Larroche, V., Paquienséguy, V. (2018). Cadres d'usage des données par des développeurs, des data scientists et des data journalistes. Livrable n°3. [Rapport de recherche] EA 4147 Elico, https://hal.archives-ouvertes.fr/hal-01730820/document.
18 https://smart-origin.com/.
19 https://www.emse.fr/~zimmermann/OpenSensingCity/peniche.html.

Figure 2. Capture d'écran de la page web dédiée à l'expérimentation.

Expérimentation OpenSensingCity, La Péniche, Grenoble le 13 décembre 2017

ARMINES MINES Saint-Étienne elico Équipe de recherche de Lyon en sciences de l'information et de la communication antidot HIKOB LEARN FROM REALITY.

L'objectif de cette journée est de concevoir une extension d'une application existante en exploitant des données ouvertes ainsi que des outils du Web sémantique, dont certains issus du projet de recherche OpenSensingCity.

L'application existante, réalisée par nos soins, permet de trouver des parkings proches avec leur disponibilité en nombre de places libres, à la manière de l'application Gre.park. En revanche, si l'application Gre.park ne fonctionne que pour les parkings de Grenoble, notre application peut s'adapter à tout territoire dès lors qu'un jeu de données ouvertes est identifiable localement.

Cette page est un point d'entrée vers la documentation des divers outils mis en œuvre pour l'expérimentation :

- descriptif des étapes de travail de l'atelier : cette page décrit ce qui est attendu des participants durant cette journée.
- Préliminaires sur quelques technologies du Web sémantique : une introduction allégée pour les participants n'étant pas familier des fondements des technologies Web sémantique.
- documentation du service de recherche de jeux de données d'Antidot : l'entreprise Antidot offre des outils permettant l'extraction, l'indexation et la recherche dans des données plus ou moins structurées. Cette page documente l'utilisation d'une API web de recherche d'information personnalisée.
- Site Web du langage de transformation de données SPARQL-Generate : ce site contient la documention du langage ainsi qu'une interface de test en ligne, le code source, un outil en ligne de commande.
- L'application Parking-loader : cette page est une application Web permettant le chargement de jeux de données parking avec leur disponibilité temps réel. La distance par rapport à un point de référence permet de trouver le plus proche avec des places libres.
- Le service Web parking et trafic d'Hikob simule des données de disponibilité de places de stationnement autour de la Péniche, ainsi que des informations sur le trafic.

Le scénario proposé supposait la recherche des jeux de données *via* le moteur de recherche de l'outil Antidot et le service web parking et trafic de la société Hikob, l'exploration et la transformation des données dans un format pivot grâce à SPARQL-Generate et l'augmentation des fonctionnalités d'une application mise à disposition.

D'abord, Antidot a mis à disposition des participants un prototype de son service de recherche de jeux de données Antidot OSC Search Box[20]. Il s'agit d'un ensemble d'outils permettant l'extraction, l'indexation et la recherche dans des données plus ou moins structurées. Si l'interface web permet de se faire une idée des données, une API web assure une recherche d'information personnalisée[21]. L'équipe Elico a participé au développement de ce service de recherche de jeux de données en réalisant une taxonomie/un mapping permettant de réconcilier les mots-clés utilisés pour accéder aux données ouvertes analogues issues de différents portails. Ce sont principalement les données ouvertes des portails métropolitains utilisant les catalogues DCAT qui ont été indexées sur le portail.

20 http://eval02.partners.antidot.net/default/osc/portal/#/.
21 http://eval02.partners.antidot.net/default/osc/portal/Antidot_OSC_UseCase.html.

Ensuite, l'entreprise Hikob a proposé aux participants un service Web parking et trafic qui simulait des données de disponibilité de places de stationnement en temps réel dans un quartier de Grenoble, autour de la Péniche, ainsi que des informations sur le trafic.

Une fois les données nécessaires identifiées et récupérées, les participants étaient orientés vers l'utilisation de l'outil SPARQL-Generate, conçu par les chercheurs de l'EMSE[22]. Extension de SPARQL 1.1, l'outil permet d'interroger des jeux de données RDF et des documents dans des formats arbitraires. Dans le cadre de l'expérimentation, les participants devaient utiliser SPARQL-Generate comme un outil de transformation des données ouvertes en modèle RDF. Pour cela, ils avaient à leur disposition une interface de test en ligne, le code source et un outil en ligne de commande.

Les outils Antidot et SPARQL-Generate ont été accompagnés d'une documentation importante et de tutoriels. Les inscrits ont reçu des informations sur le déroulement de la journée et les liens vers la documentation et les tutoriels quelques jours avant l'expérimentation. En outre, une page de présentation des technologies du web sémantique a été mise à leur disposition.

Enfin, les chercheurs de l'EMSE ont conçu une application web Parking-loader[23] qui permet de trouver des parkings proches avec leur disponibilité en nombre de places libres, à la manière de l'application Gre.park[24]. En revanche, si l'application Gre.park ne fonctionne que pour les parkings de Grenoble, l'application proposée peut s'adapter à tout territoire dès lors qu'un jeu de données ouvertes est identifiable localement et est chargé dans le bon format. En effet, l'application fonctionne uniquement avec les données en RDF (JSON-LD). L'application permet de croiser les jeux de données statiques (adresse parking, information sur le parking) avec les jeux de données dynamiques (disponibilité en temps réel). Le calcul de la distance par rapport à un point de référence permet de trouver le parking le plus proche avec des places libres.

22 https://ci.mines-stetienne.fr/sparql-generate/.
23 http://opensensingcity.emse.fr/parkingloader/#/.
24 https://play.google.com/store/apps/details?id=md.org.grepark&hl=fr.

Scénario

Chaque groupe devait suivre les trois tâches de chaîne de traitement de données avec l'ensemble des outils mis à disposition pour arriver *in fine à* augmenter l'application fournie (Parking-loader) avec de nouvelles fonctionnalités croisant plusieurs jeux de données.

Tâche 1 : Récupérer des données compatibles automatiquement

Comme les jeux de données utilisables par l'application Parking-loader sont en format JSON-LD, une première étape consistait à découvrir des jeux de données sur les parkings. Cela pouvait se faire de façon automatique en utilisant l'outil d'indexation Antidot OSC Search Box. L'objectif de cette tâche était de faire en sorte que l'application soit capable de retrouver d'elle-même des jeux de données adaptés à ses besoins. Dans ce cas, il s'agissait de trouver des données parking dans un certain format.

Tâche 2 : Transformer des données ou utiliser une transformation existante

Les jeux de données ouvertes sont généralement très hétérogènes dans leur format, dans leur structure et dans leur contenu (cf. Livrable 2[25]). Ainsi, même avec l'outil de recherche d'Antidot, certains jeux de données clairement identifiés comme « parking temps réel » ne pouvaient pas être utilisés directement par l'application. Toutefois, l'outil d'Antidot indique également s'il existe une transformation exécutable permettant de passer du format initial d'un jeu de données vers le format pivot (JSON-LD) que consommait l'application proposée. Il s'agissait alors de retrouver des données parking puis d'appliquer la transformation.

Parfois, aucune transformation n'était fournie. Il s'agissait alors pour les participants de construire par eux-mêmes une ou des transformations leur permettant de consommer un ou des jeux de données supplémentaires. Les participants pouvaient réaliser des transformations de deux façons : soit en utilisant le langage et les outils qu'ils connaissaient, soit en utilisant le langage et le moteur de transformation SPARQL-Generate.

Tâche 3 : Étendre les fonctionnalités de l'application en croisant d'autres jeux de données

25 Paquienséguy, F., Dymytrova, V. (2017). Analyse de portails métropolitains de données ouvertes à l'échelle internationale. Livrable 2. [Rapport de recherche]. Elico (Équipe d'accueil lyonnaise en Sciences de l'information et de la communication). <hal-01449348>.

Pour cette tâche, des données trafic simulées par Hikob dans un quartier de Grenoble ont été fournies en plus des données parking. Les données trafic ont été formatées de telle sorte que des liens explicites étaient faits entre celles-ci et les données parking. En particulier, des identificateurs de référence ont été utilisés pour les rues et les adresses. Il s'agissait de tenir compte de la densité de trafic pour présenter ou suggérer des parkings ou places de stationnement disponibles.

Cette tâche était basée sur une projection dans l'avenir et une supposition que de nombreuses données en lien avec le trafic et le transport seront à l'avenir disponibles en open data sur de nombreux territoires. Le but de cette tâche ouverte était de concevoir un projet d'application exploitant plusieurs jeux de données pouvant exister sur plusieurs territoires urbains. Les données pressenties pour concevoir ces fonctionnalités peuvent être fictives avec une projection vers l'avenir où elles seraient rendues un jour disponibles en open data.

Matériaux collectés

Répartis en trois groupes, les participants ont été invités à restituer les résultats de leur travail en fonction de plusieurs éléments : les données utilisées, les étapes essentielles de la réalisation, les outils utilisés, les points de négociation dans le groupe et l'avis sur les données et les outils mis à disposition. Par ailleurs, nous avons demandé à chaque groupe d'accompagner leur restitution de captures d'écran jugées significatives.

À l'issue de la journée, nous avons constitué un corpus multimodal et multi source des données :

- Photographies prises lors de la journée (organisation de l'espace du challenge et dispositions des participants) ;
- Enregistrements audio des interactions entre les participants de chaque groupe ;
- Vidéos des étapes clés de la journée (présentation des outils et des scénarios, interactions entre les concepteurs et les participants, restitutions des résultats) ;
- Notes d'observation.

Dans les jours suivant l'événement, nous avons effectué les entretiens post expérimentation d'une heure environ avec deux participants de l'expérimentation.

3. ANALYSES DES EXPÉRIMENTATIONS À TRAVERS LA CHAÎNE DU TRAITEMENT DES DONNÉES

Nous présenterons ici les analyses des expérimentations en fonction de la chaîne du traitement des données identifiées dans le livrable 3, réalisé par Elico[26]. En effet, dans ce document, nous avons souligné qu'au-delà des caractéristiques professionnelles, tous les réutilisateurs des données ouvertes sont confrontés à la chaîne de traitement comprenant la collecte et l'exploration, la compréhension et l'analyse des données, la transformation et enfin, l'exploitation ou l'implémentation des données à travers un développement ou une modélisation. Nous allons voir dans cette partie comment les outils développés dans le cadre de notre ANR s'inscrivent dans les trois étapes centrales de la chaîne de traitement de données par des professionnels du web, ayant participé aux deux expérimentations. Nous allons nous focaliser ainsi sur la récupération, le traitement et l'implémentation des données.

3. 1. MÉTHODOLOGIE

Lors des deux expérimentations, les chercheurs en informatique se sont concentrés sur le fonctionnement technique des outils. Leur objectif était la « concrétisation »[27] des outils, au sens de G. Simondon. Ils ont donc mobilisé les critères d'évaluation empruntés aux méthodes du domaine des « interfaces homme-machine » : bon comportement technique, rapidité de l'utilisation et faible taux d'erreur dans la manipulation des outils[28]. Ils ont donc essayé de repérer les bugs techniques et de les résoudre.

Nous, les chercheurs en SIC, avons davantage prêté l'attention à la situation de communication particulière que les tests des outils engagent en analysant les interactions entre les participants et les outils mais aussi les interactions au sein des groupes de

26 Dymytrova, V., Larroche, V., Paquienséguy, V. (2018). Cadres d'usage des données par des développeurs, des data scientists et des data journalistes. Livrable n°3. [Rapport de recherche] EA 4147 Elico, https://hal.archives-ouvertes.fr/hal-01730820/document.
27 Simondon, Gilbert (1958), Du mode d'existence des objets techniques, Paris, Aubier, p. 47.
28 Nielsen, Jacob (1993), Usability engineering. Boston, Academic Press.

participants. Notre objectif était d'évaluer le potentiel de l'appropriation socioprofessionnelle de ces outils. Nous avons donc considéré les participants moins comme testeurs des outils mais comme contributeurs à leur construction et à leur finalisation. En effet, il semble difficile d'imaginer que les participants puissent avoir des attentes ou des besoins vis-à-vis d'une technologique et des outils qu'ils ne connaissent pas et dont les potentialités sont testées. En revanche, les participants pouvaient formuler de propositions concernant les fonctionnalités et les usages possibles des outils. Nous nous sommes donc focalisés sur le repérage et l'analyse des stratégies signifiantes des participants. L'hypothèse principale était la suivante : dans la situation d'expérimentation, les participants adapteront leurs cadres de référence aux modes d'action et aux outils proposés pour réaliser le scénario proposé. Les chercheurs en SIC se positionnent ici comme parties prenantes de l'ingénierie du dispositif expérimental.

3.2. RECHERCHE ET RÉCUPÉRATION DES DONNÉES

Quel que soit le scénario proposé, au Tubà et à La Péniche, les participants des deux expérimentations ont été d'abord confrontés à la récupération des données. Chronophage dans les deux cas, cette étape consistait pour Tubà dans la récupération des données dans le format JSON-LD à partir de la page dédiée à l'expérimentation, car la plateforme OSC Search Box Antidot n'était pas encore au point. Dans le cas de l'expérimentation à La Péniche, les participants accédaient aux données à partir d'un point d'entrée unique que constitue OSC Search Box Antidot et avaient la possibilité de récupérer les données dans le format de leur choix, JSON-LD ou autre.

La récupération des données au Tubà

Non habitués au format JSON-LD dans lequel les données ont été mises à disposition, les participants ont mis du temps à les récupérer et à les intégrer dans leurs espaces de stockage et dans leurs logiciels de traitement et d'analyse de données.

En effet, plusieurs participants au profil data scientist et data analystes ont souligné lors des entretiens post expérimentation que le format JSON-LD n'était pas compatible avec plusieurs solutions du marché qu'ils utilisent et notamment avec le logiciel R, leur outil de prédilection : « *On a passé énormément de temps pour récupérer les données. Le format*

n'est pas idéal c'est du JSON-LD qui n'est pas très optimisé. Ce n'est pas assez packagé. On utilise en général des choses déjà faites, on ne code pas l'algorithme de base à chaque fois. On utilise des packages : pour charger un CSV, pour lire un JSON, ce sont des choses qui ont déjà été faites et qui sont disponibles dans des librairies. Les librairies qu'on a utilisées n'arrivaient pas à interpréter le format » (participant data analyste, entretien post expérimentation).

De fait, pour récupérer les données en JSON-LD et les restituer dans un espace de stockage ou les intégrer dans un outil de traitement de données, les participants ont développé des stratégies de contournement : « *Par hasard, j'ai trouvé une astuce : appeler le portail depuis le navigateur, copier-coller la réponse dans un txt et après de charger ce txt sous R et là, il arrivait à l'interpréter* » (participant data analyste lors de l'entretien post expérimentation). D'autres participants ont contourné les problèmes liés à la récupération des données en format JSON-LD en allant chercher les données directement sur les portails dédiés.

En revanche, les participants-développeurs semblent avoir réussi à s'approprier plus facilement des données dans le format JSON-LD, même s'ils n'étaient pas très familiers des technologies du web sémantique. Ils ont sollicité les chercheurs de l'EMSE pour écrire certaines requêtes : « *Après, le langage n'est pas très compliqué quand on a un exemple* » (participant développeur lors de l'entretien post expérimentation). Cela témoigne de la prégnance des pratiques professionnelles des acteurs, ayant différentes capacités d'adaptation à des formats multiples de données existantes.

Malgré un travail important de la préparation des données par les chercheurs de l'EMSE, les participants ont rencontré plusieurs difficultés avec la récupération. Par exemple, en début de l'expérimentation, les données des parkings de Lyon n'ont pas été accessibles à cause d'un problème d'authentification. Ces problèmes d'authentification semblent assez difficiles à résoudre de manière automatique et générale, chaque ville ayant, pour certaines données, son propre protocole authentification. Un autre problème soulevé par les participants concernait l'absence de cohérence et de normalisation des champs de données ouvertes mises à disposition ce qui constituait un problème pour leur intégration dans une application ayant pour vocation de couvrir plusieurs territoires. L'outil SPARQL-Generate a pour vocation de répondre à cette problématique.

La récupération des données à La Péniche

Lors de la deuxième expérimentation, les participants n'ont pas consacré beaucoup de temps à la récupération des données. Ils ont utilisé le portail Antidot OSC Search Box d'une manière exploratoire en faisant les recherches par mots-clés et par facette. Une fois un ou deux jeux de données récupérés, ils se sont concentrés sur l'utilisation de l'outil SPARQL-Generate : « *L'API d'Antidot, on a pu la tester, mais mon but ce n'était pas vraiment d'apprendre à récupérer les données par ce moyen-là. J'avais plus envie de voir les sources de données et comment on pouvait les transformer* » (participant ingénieur R&D, entretien post expérimentation). Le fait que le portail Antidot OSC Search Box a été sous-utilisé s'explique ainsi en partie par des préoccupations professionnelles des participants dont plusieurs ayant le profil data analystes ou data scientists, peu confrontés à la recherche des données car travaillant beaucoup avec des données fournies par des clients.

La présence d'un scénario à suivre a fortement réduit la nécessité d'utiliser le portail pour rechercher et collecter une grande variété de données. Les participants ont fait l'économie du temps et des efforts nécessaires à l'exploration des fonctionnalités du portail, la consultation de sa documentation et son utilisation pour la construction des requêtes. Le nombre de données dont ils avaient besoin leur permettait de contourner l'utilisation du portail. Certains participants ont utilisé directement les portails des données ouvertes. Par exemple, un groupe s'est servi de l'API de Métro Mobilité de Grenoble qu'ils connaissaient avant d'autant plus que les parkings de Grenoble n'étaient pas répertoriés dans le portail d'Antidot. Cela s'explique par la facilité d'indexation des données ouvertes présentées dans les catalogues DCAT, ce qui n'est pas le cas du portail métropolitain de l'open data de Grenoble.

Si d'une manière générale, les participants reconnaissent l'importance d'un outil permettant de localiser les données ouvertes – et hétérogènes - situées dans une multiplicité d'endroits, certains restent perplexes sur la pérennité d'un tel service : « *Cela amène en effet une sorte de standardisation mais j'ai du mal à voir comment cela peut être réutilisé par la suite. Il faut des ressources importantes, compte tenu du temps de travail important que nous consacrons pour ramener nos données à l'intérieur de la collectivité vers une sorte de standardisation* » (participant chargé de mission open data, entretien post expérimentation).

Synthèse

Le livrable 3 consacré aux cadres d'usage des données par des professionnels des data a déjà souligné le fait que le choix des jeux de données et leur récupération sont étroitement liés à des objectifs visés et des pratiques socioprofessionnelles des acteurs. Les deux expérimentations menées confirment ce constat. En effet, lors de deux événements les participants se sont retrouvés face à des scénarii et des objectifs fixés par les organisateurs qui supposaient la recherche et la récupération des jeux de données liées à la mobilité (parking, trafic, météo, etc.) à l'échelle de plusieurs métropoles françaises. Peu habitués de ce type de croisement de données multi-territoires, les participants se sont souvent limités à une ville ce qui ne permettait pas de réellement évaluer l'efficacité de la recherche et de la récupération des données à partir du portail Antidot OSC Search Box. La présence d'un scénario à suivre a réduit le besoin d'utiliser le portail pour rechercher et collecter une grande variété de données. En même temps, plusieurs participants ont fait le choix d'aller récupérer des données ouvertes directement sur les plateformes métropolitaines dédiées qu'ils connaissaient avant l'expérimentation.

Ensuite, les pratiques socioprofessionnelles des participants ont fortement impacté la manière dont ils ont envisagé la récupération des données et leur intégration. L'expérimentation au Tubà supposait la récupération des données par le téléchargement, ce qui pourrait moins convenir aux professionnels habitués des web services et des API. A La Péniche, à partir du portail Antidot OSC Search Box, les participants pouvaient aussi bien télécharger les jeux de données que faire l'usage de l'API. Dans les deux cas, nous avons constaté que le langage RDF nécessite un temps d'apprentissage et de prise en main. C'est aussi le cas du format de données JSON-LD qui par ailleurs n'est pas compatible avec les outils professionnels des réutilisateurs. Si les développeurs se sont mieux retrouvés dans JSON-LD que les data scientists, car ils avaient l'habitude d'utiliser le format JSON, ils ont été surpris par la structure des données en graphe RDF. Les participants des deux expérimentations quel que soit leur profil n'étaient pas suffisamment outillés en compétences pour manipuler les données du web sémantique.

Enfin, plusieurs problèmes liés à la qualité des données ouvertes mises à disposition (par

ex., champs manquants, absence de cohérence des champs, identification nécessaire, etc.) ont été soulevés par les participants qui s'attendaient lors de l'expérimentation à avoir des données réconciliées et nettoyées.

3.3. TRANSFORMATION DES DONNÉES

Dans la première expérimentation au Tubà, la transformation des données en modèle RDF et au format JSON-LD a été réalisée au préalable par les chercheurs de l'EMSE. Les participants devaient toutefois comprendre la structuration des jeux de données pour pouvoir se les approprier et intégrer ces données dans leurs outils de traitement ou d'analyse. L'outil de transformation SPARQL-Generate ne faisait pas l'objet du test au Tubà. En revanche, à La Péniche, les participants devaient utiliser SPARQL-Generate comme un outil de transformation des données ouvertes en modèle RDF. Pour cela, ils avaient à leur disposition une interface de test en ligne, le code source et un outil en ligne de commande.

SPARQL-Generate au Tubà

Les participants de l'expérimentation ont passé beaucoup de temps à comprendre la structure des données en RDF. Peu familiers des technologies du web sémantique, ils ne pouvaient pas se lancer rapidement dans la manipulation des données et leur intégration. Les uns ont sollicité de l'aide de la part des concepteurs de SPARQL-Generate, présents lors de l'expérimentation pour extraire les données (statiques et dynamiques) en l'état puis s'atteler au calcul des distances, afin de lier l'ensemble dans le cadre de leur application. D'autres ont préféré travailler avec les formats et les outils qui leur étaient familiers : « L*es fichiers ne me donnaient pas envie de rentrer là-dedans* » (un data analyste, entretien post expérimentation). Ils ont donc passé beaucoup de temps à transformer les données dans les formats compatibles avec leurs solutions professionnelles (logiciel R). Enfin, un groupe a adopté une approche beaucoup plus compréhensive en ayant une posture immersive dans les données et l'outil qui les a générées. Ils ont en effet choisi de travailler directement depuis l'outil SPARQL-Generate afin de mieux intégrer son fonctionnement et d'enrichir le contenu par la création d'un nouveau champ, susceptible une fois les données extraites de contribuer à une application "en temps réel" (place de parking disponible à l'instant T). Ce groupe se distinguait des autres par des connaissances préliminaires des technologies du

web sémantique–d'un des participants, qui a donc pu jouer le rôle de « passeur » de ces connaissances envers d'autres membres du groupe. Comme ils ont choisi de travailler à l'intérieur de SPARQL-Generate, ils n'ont pas donc été confrontés aux transformations des formats.

D'une manière générale, les participants ont souligné l'importance de la documentation et des tutoriels pour comprendre comment les données ont été générées et structurées et pouvoir s'approprier les données structurées en RDF et manipuler le format JSON-LD. Par ailleurs, plusieurs participants au profil développeur ont reconnu que la même structuration des données assurée par les transformations avec SPARQL-Generate est susceptible de faciliter la manipulation des données aux formats et aux structurations hétérogènes : « *C'est ça, l'avantage. C'est que si vous normalisez toutes les villes avec une écriture en RDF type et avec une utilisation d'ontologies, notre algorithme va fonctionner pour toutes les villes…*» (Développeur, entretien post expérimentation).

SPARQL-Generate à La Péniche

A La Péniche, les participants se sont principalement investis dans l'exploration et l'utilisation de SPARQL-Generate. Malgré le souhait de réunir les développeurs spécialisés en web sémantique, nous nous sommes retrouvés avec un public varié et ayant peu de connaissance dans la matière. Le manque de compétences en web sémantique s'est senti dans la dynamique d'avancement des groupes. En effet, seuls deux participants avaient des connaissances préalables du modèle RDF et du langage Sparkl dont SPARQL-Generate est le dérivé. Ces participants ont toutefois mis du temps à s'approprier l'outil et ont dû passer par plusieurs essais de manipulation et erreurs avant d'arriver au résultat souhaité. D'autres, sans connaissances en web sémantique, se sont concentrés sur la compréhension du langage RDF et de l'outil SPARQL-Generate en étudiant la documentation et les tutoriels. A la différence de Tubà, lors de cette deuxième expérimentation les participants n'ont pas mobilisé leurs propres outils pour contourner le web sémantique et l'étape de la transformation des données. Ils ont travaillé les jeux de données statiques et dynamiques liées aux parkings avec SPARQL-Generate. Compte tenu du temps important de l'appropriation de cet outil, ils n'ont pas eu le temps de suivre le scénario proposé jusqu'au bout et d'augmenter les fonctionnalités de l'application Parking-loader mise à disposition.

Synthèse

Si plusieurs participants des deux expérimentations reconnaissent l'importance d'avoir un outil de transformation et d'homogénéisation des formats et des structures de données en RDF, la plupart d'entre eux soulignent la nécessité d'un réel accompagnement dans la prise en main de SPARQL-Generate. Quasiment absente lors de la première expérimentation au Tubà, une documentation et des tutoriels concernant cet outil ont été développés pour les participants de l'expérimentation à La Péniche. Toutefois, la consultation de la documentation et la formation par le tutoriel nécessitent un temps considérable ce qui est difficilement conciliable avec les objectifs d'une expérimentation organisée autour de la réalisation d'un scénario précis. Dans ce cas-là, il est important d'avoir des publics ayant déjà des bonnes compétences en web sémantique car il ne s'agit pas d'un outil clé en mains destiné à un professionnel du web lambda. SPARQL-Generate est davantage destiné à un bon connaisseur du web sémantique.

3.4. RÉUTILISATIONS

Les dynamiques des deux expérimentations ont été différentes. En effet, au Tubà les participants ont tout de suite imaginé les façons dont ils pourront implémenter les données ouvertes proposées. Ils se sont beaucoup appuyés sur leurs propres outils professionnels, ce qui leur a permis de faire l'économie du temps et d'arriver à réaliser dans un temps imparti des productions plus ou moins fonctionnelles.

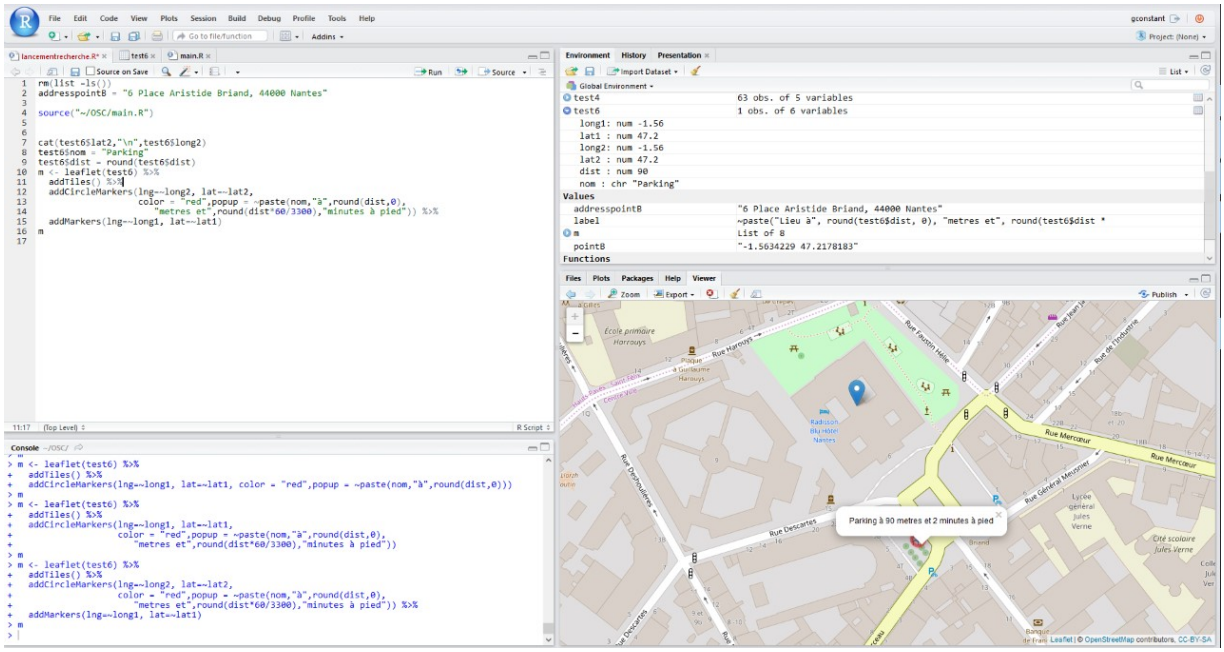
A La Péniche, les participants ont fait le choix de suivre la trame proposée par des organisateurs et ont manipulé successivement deux outils : Antidot OSC Search Box et SPARQL-Generate. De fait, ils n'ont pas eu le temps d'arriver à la troisième étape proposée, à savoir l'augmentation des fonctions de l'application Parking-loader. Certains participants ont soulevé dans les entretiens post expérimentation la question d'une faible problématisation du scénario par des organisateurs ce qui a constitué pour eux un élément de blocage. Ils s'attendaient à avoir un problème plus précis à résoudre dans le cadre du scénario proposé.

Focus sur les réalisations des participants au Tubà

Tout au long de l'expérimentation au Tubà, les participants ont travaillé en 4 groupes. La plupart des groupes ont gardé à l'esprit le scénario initial, et en ont donc respecté les grandes étapes mais avec des modalités de travail fort différentes.

Ainsi, le groupe 1 a réuni un data ingénieur, un data scientiste et un développeur ce qui leur a permis de distribuer les tâches en fonction des compétences de chacun. Le data ingénieur s'est occupé de la récupération et de l'intégration des données, le data scientiste a réalisé les analyses et la modélisation statistique alors que le développeur a travaillé sur l'interface web. Ce groupe a créé un calculateur d'itinéraire sous forme de cartographie permettant de trouver un parking le plus proche du lieu de destination (Fig. 4, ci-dessous).

Figure 4. Calculateur d'itinéraire du parking le plus proche (groupe 1).



L'application a été réalisée à partir des données ouvertes des parkings de Nantes, Strasbourg et Rennes (noms des parkings, adresses, géolocalisation, nombre de places) et avec leurs outils professionnels habituels. Ils ont notamment utilisés :

- MongoDB, un système de gestion des bases de données ;

- Talend, un logiciel libre français de type ETL[29], qui permet d'interagir avec les données sans coder en créant graphiquement des processus de manipulation et de transformation de données puis en générant l'exécutable correspondant sous forme de programme Java ou Perl ;
- Logiciel R, pour l'analyse statistique.

Figure 5. Recherche de stationnement le plus proche (groupe 2).

```
def test_premier():
    staticPage = 'http://opensensingcity.emse.fr/tuba/data/static-rdf/nantes/nantes.parking.ttl'
    dynamicPage = 'http://opensensingcity.emse.fr/sparql-generate/api/transform?queryurl=http%3A%2F%2Fopensensingcity.emse.fr%2
    pk = Namespace("http://opensensingcity.emse.fr/ontologies/parking/")

    graphStatic=rdflib.Graph()
    graphStatic.load(staticPage, format='n3');

    graphDynamic=rdflib.Graph()
    graphDynamic.load(dynamicPage);

    graph = rdflib.Graph()
    graph = graphDynamic + graphStatic

    #geolocator = geopy.geocoders.Nominatim()
    #location = geolocator.geocode("1 Place Charles Beraudier, 69003 Lyon, France")
    #print((location.latitude, location.longitude))

    for s, p, o in graph.triples((None, RDF.type, pk.ParkingPlace)):
        for p,o in graph.predicate_objects(s):
            if p == rdflib.URIRef('http://www.w3.org/2000/01/rdf-schema#label'):
                print "label: " + o
            if p == rdflib.URIRef('http://www.w3.org/2003/01/geo/wgs84_pos#long'):
                long = o
                print "long: " + o
            if p == rdflib.URIRef('http://www.w3.org/2003/01/geo/wgs84_pos#lat'):
                print "lat: "+o
                lat = o

        break
```

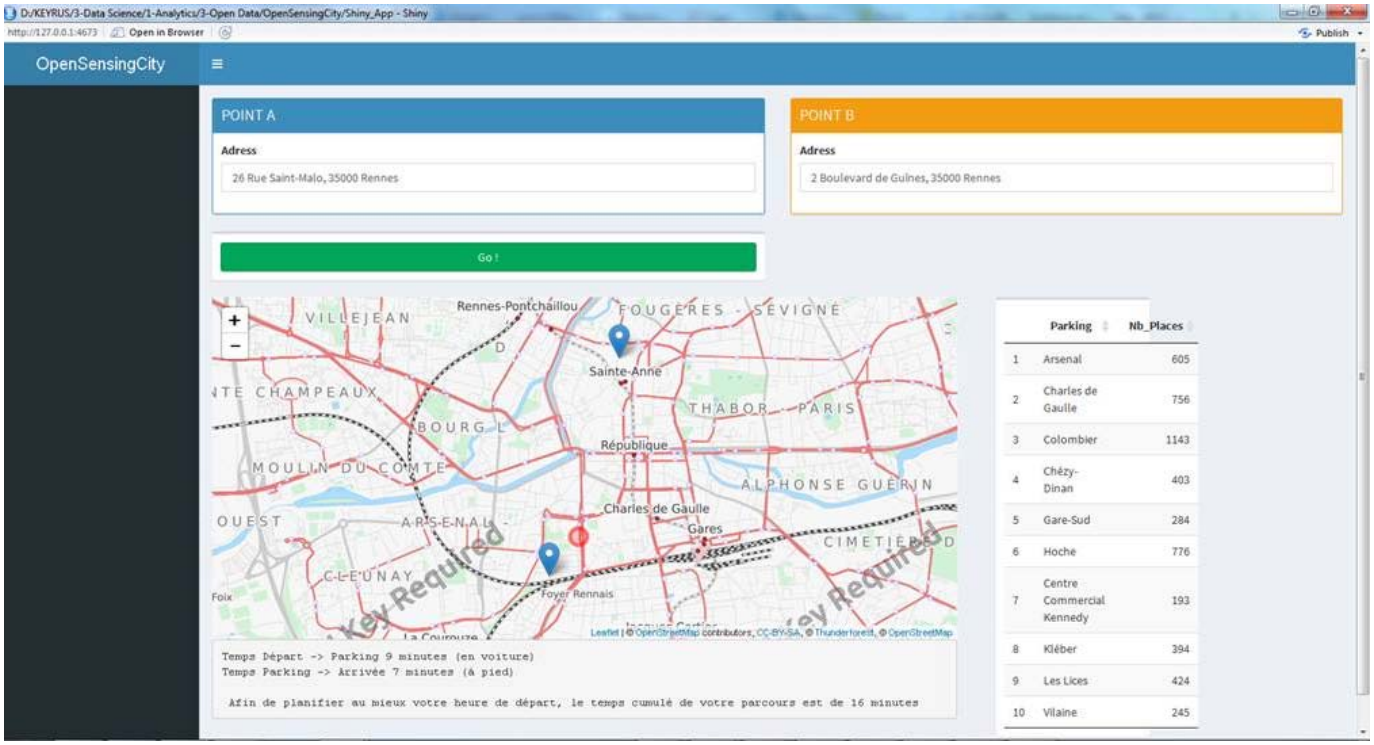
Grâce à une configuration très proche du fonctionnement des entreprises du numérique où il existe une collaboration étroite entre les métiers des données et du web, ce groupe a réussi à réaliser rapidement le scénario.

Le groupe 2 a réuni un doctorant et un étudiant en informatique de l'INSA de Lyon. Ayant suivi quelques heures de cours en web sémantique, ils ont fait le choix de suivre au plus près

29 L'ETL (Extract, Transform, Load) est un processus d'intégration des données qui permet de transférer des données brutes d'un système source, de les préparer pour une utilisation en aval et de les envoyer vers une base de données, un entrepôt de données ou un serveur cible. Dans ce processus la transformation des données intervient sur un serveur intermédiaire avant le chargement sur la cible. Source : https://www.lemagit.fr/definition/ETL-et-ELT.

les consignes données et de proposer une solution générique de recherche de stationnement le plus proche, basée sur l'utilisation des graphes RDF (Fig. 5, p. 24). Pour cela, ils ont utilisé des données des parkings mises à disposition pour Nantes, Paris et Strasbourg. Ils ont entre autres mobilisé le langage de programmation Python (GeoPy), l'API de Google Map et l'application Postman, proposé par Google Chrome pour faire des requêtes dans les API et les tester. Encore une fois, nous constatons ici la prégnance des outils familiers aux participants.

Figure 6. Interface web de l'application parking (groupe 3).



Le groupe 3 a réuni deux data analystes d'une société spécialisée en traitement des données. Fidèles à leurs habitudes professionnelles, ils se sont focalisés sur le produit en décidant de produire l'application le plus rapidement possible et l'améliorer par la suite. Celle-ci devait afficher le parking le plus proche sur un parcours entre A et B (calcul de distance) et permettre à l'usager de choisir son point d'arrivée et son point de départ en direct. Pour cela, ils ont utilisé les outils qu'ils connaissaient bien : le logiciel R et la librairie Shiny qui fonctionne avec celle-ci et qui permet de produire une application web assez rapidement. Pour faire circuler les données entre les participants, ils se sont appuyés sur le

modèle de versionning propre au GitLab, qui permet de travailler simultanément sur les données.

La formation et les origines métier des participants ont joué un rôle important dans le respect des consignes et la façon dont ils se sont appropriés les outils et les données. Les uns ont en effet privilégié la satisfaction rapide de la commande (l'application) quitte à revenir ensuite sur les données, alors que d'autres ont choisi d'abord d'étudier l'architecture des outils proposés et des données pour ensuite pouvoir les enrichir de façon plus efficace en amont même de la création d'une application. Les postures exploratoires sont donc fortement contraintes par les habitus professionnels.

3.5. MISE EN PERSPECTIVE DES OUTILS

Les deux expérimentations ont permis d'observer les usages des outils OSC développés dans notre ANR par divers professionnels du web et des données et d'apporter des améliorations importantes à ces outils à la fois pendant et après les expérimentations.

OSC Search Box d'Antidot

Lors de la première expérimentation, le portail d'indexation des jeux de données ouvertes OSC Search Box conçu par Antidot n'était pas encore au point. Il référençait un nombre de jeux de données limité et aucun travail de réconciliation de données n'a été entrepris à l'époque. Par ailleurs, le protocole de cette expérimentation et le scénario proposé ne supposaient pas l'usage du portail. Nous avons toutefois sollicité, lors des entretiens post expérimentation, les avis des participants sur l'utilité d'avoir un tel portail. Les retours ont été très positifs. Les interviewés ont souligné la pertinence d'un tel outil pour la recherche rapide des données concernant plusieurs territoires. Ils envisageaient un tel portail comme une forme de standardisation et d'homogénéisation des jeux de données et des métadonnées qui les accompagnent. Certains ont évoqué une éventuelle vérification des données ce qui à leurs yeux apporterait une vraie valeur ajoutée à ce portail.

C'est surtout la deuxième expérimentation, celle de La Péniche qui a permis d'observer les comportements des participants face au portail OSC Search Box d'Antidot, même si celui-ci est resté sous-utilisé, comme nous l'avons déjà souligné. Après une présentation rapide du portail par Eric Noulard, son concepteur, tous les participants sont allés l'explorer comme le

prévoyait le scénario proposé. Ils ont réalisé quelques requêtes en cherchant par mots-clés et par facettes (ville ou thématique). Toutefois, son usage ne semblait pas pleinement satisfaisant dans la mesure où les données proposées sont parfois incomplètes (par ex., absence des données parking pour Grenoble) : « *Cet outil ne répertorie pas toutes les données, si c'était un outil qui indexe tous les jeux de données ce serait génial* » (data analyste, entretien post expérimentation). En même temps, contraints par un scénario d'usage, les participants n'avaient pas non plus besoin d'utiliser pleinement le potentiel de l'outil qui est surtout intéressant pour la recherche et la récupération d'un grand nombre de données multi-territoires. Enfin, nous avons déjà souligné une interrogation sur la pérennité d'un tel service, exprimé par un chargé de mission open data participant à l'une des expérimentations.

SPARQL-Generate

Cet outil a tiré un maximum des deux expérimentations car il y a été central. Les retours des participants sont mitigés. Plusieurs, en particulier à La Péniche, ont été séduits par la promesse de SPARQL-Generate, à savoir avoir une même structuration et un même format des données issues des métiers différents. Certains ont également apprécié l'interface proposée.

Toutefois, la prise en main de SPARQL-Generate nécessite un véritable accompagnement. Tous les participants ont insisté sur l'importance de la documentation et des tutoriels concernant SPARQL-Generate et le langage RDF lui aussi peu connu par des publics réunis au Tubà et à La Péniche. Autrement, l'usage de l'outil nécessiterait des intermédiaires, familiers des technologies du web sémantique. Si à la Péniche une documentation et des tutoriels cette fois très détaillés ont été mis à disposition des participants quelques jours avant la deuxième expérimentation, les participants se sont sentis désarçonnés par le volume important de nouvelles choses à apprendre. Une nouvelle fois, cela permet de faire le constat que des participants - non familiers d'un outil et/ou d'un format - adoptent une attitude assez proche lorsqu'ils sont confrontés à un exercice en temps limité : mobiliser des outils/formats issus de leur environnement professionnel. La documentation fournie ne les incite donc pas davantage à faire usage de l'outil, car la contrainte temporelle et le principe de réalité (achever le scénario de départ) dominent. Ils contournent alors ces écueils en

utilisant des outils/formats connus d'eux, ce qui leur permet de gagner du temps et de passer ensuite à la tâche suivante.

L'accessibilité du format JSON-LD reste aussi problématique pour certains participants en raison de son incompatibilité avec des solutions professionnelles existantes comme par exemple le logiciel R. D'autres soulignent qu'il s'agit d'un outil en phase de développement qui nécessite du temps pour se faire connaître et s'affirmer : « *Ce qui me préoccupe c'est d'arriver à la généralité de cet outil…Sparql a déjà une petite communauté, mais pas comme SQL qui est plus diffusé et populaire. Le risque c'est d'arriver à rendre la donnée générique, universelle mais par un outil qui n'est ni générique ni universel* » (un data scientist, entretien post expérimentation). La capacité à réunir une communauté d'usages est pour les participants interviewés un élément central dans la réussite de SPARQL-Generate : « *Est-ce que les outils vont s'adapter à tout ce qui est JSON-LD, je ne sais pas trop. C'est la masse qui va décider. Soit les gens vont comprendre que JSON-LD a un intérêt, du coup les outils vont migrer, soit les gens vont développer leurs outils et si la majorité des outils n'utilise pas JSON-LD, le langage meurt. Globalement, c'est comme ça que cela se passe. C'est la masse, c'est la communauté qui décide : si elle dit « l'outil est bien et je reste sur l'outil et tant pis j'utilise un autre format, soit elle dit le format est vraiment bien et ce sont les outils qui vont changer »* (un data ingénieur, entretien post expérimentation).

Les participants des expérimentations ont surtout rencontré les difficultés liées à la compréhension du langage RDF, fondement du Web sémantique. Comme le confirme l'un des participants : « *le souci ce n'est pas les données, c'est le langage utilisé* » (un développeur, entretien post expérimentation). En effet, pour plusieurs participants, il s'agit d'un nouveau paradigme des données où les relations entre les données sont très différentes, conditionnées par un formalisme logique et non pas par les relations propres aux bases de données relationnelles. Si certains ont entendu parler du web sémantique et même ont suivi quelques cours, ils n'ont pas eu l'occasion de le pratiquer : « *Cela m'a ouvert l'esprit sur le concept, les outils, mais cela reste flou par rapport aux façons dont on pouvait le mettre en œuvre* » (un participant ayant suivi le MOOC sur le web sémantique).

Comme l'ont confirmé les expérimentations, SPARQL-Generate est destiné aux spécialistes du web sémantique. La difficulté de recruter les participants au profil web sémantique en

particulier pour La Péniche constitue ainsi une vraie limite de l'expérimentation. D'une manière générale, les attendus des participants, tous professionnels des données se situent soit en amont des outils développés (pouvoir explorer la donnée pour mieux la comprendre et la modéliser selon ses besoins ensuite), soit en aval (disposer d'outils clé en main, sans passer par des phases intermédiaires de construction de la donnée).

Synthèse

Les difficultés rencontrées par les participants se focalisent principalement sur l'obligation à utiliser de nouveaux outils et à travailler à partir d'un format de données à la fois peu connu et dont le fonctionnement semble difficile à appréhender. En effet, plusieurs participants ont souligné le caractère clivant du format JSON-LD, renforçant le fait que les pratiques des métiers sont liées à des langages spécifiques parfois excluants.

La promesse initiale des outils testés, à savoir faciliter la récupération, puis la transformation des données ouvertes ne semble pas avoir suffi et ce pour trois raisons : 1) l'outil n'étant pas encore abouti à ce stade, il ne rend pas un service suffisamment efficient pour permettre de faire l'impasse sur d'autres modalités d'usage - plus familières aux participants ; 2) le service proposé - la donnée "clé en main" - détourne certains participants de leur intérêt à travailler les données à la source et à réaliser des activités d'exploration et de "bricolage", pour reprendre leurs termes, sur celles-ci. 3) Le langage et le format proposé par SPARQL-Generate - bien que les fonctions soient intéressantes - concerne en l'état une communauté trop restreinte : les développeurs, familiers du web sémantique. Cet outil serait donc à destination d'*happy few* et devrait être considéré comme un palier intermédiaire - mais nécessaire - pour faciliter la génération de données enrichies.

4. SYNTHÈSE GÉNÉRALE

Les deux expérimentations ont donc permis de dégager au moins trois stratégies signifiantes dans les usages des outils conçus par des réutilisateurs : l'impact de la chaîne de traitement des données, la prégnance des pratiques professionnelles et l'importance de la communauté d'usage.

1/ Impact de la chaîne de traitement des données

La distribution des rôles dans les groupes participant à l'expérimentation a été fortement impactée par la chaîne de traitement des données. En fonction de leur formation et de leurs compétences, les uns se sont occupés de la récupération des données, tandis que d'autres se sont investis dans la transformation, l'analyse ou l'implémentation des données. Cela amène à penser que les outils conçus, OSC Search Box et SPARQL-Generate sont complémentaires. Le premier permet d'accéder aux données aussi bien par la recherche manuelle que par le web service assuré par l'API. Le deuxième permet de transformer les données en RDF et en JSON-LD, en assurant une sorte d'homogénéisation de structure et de format des données, issues des métiers et des portails différents.

2/ Prégnance des pratiques professionnelles

Les participants se sont interrogés tous à des degrés divers sur les formats, leurs accessibilités et leurs facilités de transformation. Toutefois, chaque groupe a adopté une posture différente par rapport au choix et à la réalisation d'un scénario en mobilisant des cadres de référence professionnels connus et donc rassurants : format, logiciels, pratiques. L'ensemble "cohérent" qu'était censé proposer les outils de cette expérimentation ne semble pas, aux dires des participants, constituer un argument suffisant. Son potentiel semble avoir suscité un réel intérêt de leur part. On retrouve là les mécanismes de diffusion et de traduction (la force du réseau) mis en avant par Bruno Latour et Michel Callon (Akrich, Callon, Latour, 2006). Force est de constater à l'issue de cette expérimentation qu'en l'état actuel, ces outils, loin de générer de nouvelles formes d'agir, contribuent à renforcer les pratiques professionnelles existantes. La courbe d'apprentissage (« *learning curve* ») d'un nouvel outil, même s'il revêt des éléments familiers à des développeurs, doit prendre en compte un temps non négligeable d'appropriation et nécessite des ajustements et des «

traductions » (Akrich, Callon, Latour, 2006). Or le programme de la journée lors des deux expérimentations n'a sans doute pas laissé suffisamment de temps pour le faire.

3/ Importance de la communauté d'usage

Pour être véritablement efficient, SPARQL-Generate doit parvenir à toucher une plus grande communauté de professionnels et ne pas s'adresser qu'à un cercle restreint de spécialistes du web sémantique. Ses créateurs doivent pour cela faire un important travail de vulgarisation et de communication auprès des différentes communautés d'usagers, susceptibles de pouvoir saisir toute la richesse et la complexité d'un tel outil.