Captation, extraction et restructuration de données À partir de sources numériques hétérogènes

Violaine Guichet, Mathilde Plard

CAPTATION, EXTRACTION ET RESTRUCTURATION DE DONNÉES À PARTIR DE SOURCES NUMÉRIQUES HÉTÉROGÈNES

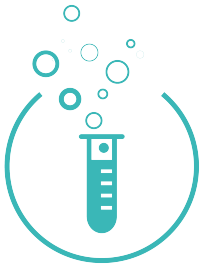
AUTEURES

Violaine GUICHET (IE), Mathilde PLARD (CNRS, UMR 6590 ESO)

RESUME

L'article revient sur les étapes de conception et de construction du jeu de données structurées et harmonisées dans le cadre du projet RUNNING DATALAB[1].

Sur le format d'un data paper il s'agit de présenter ici, chemin faisant, l'avancée du travail réalisé. D'autres versions viendront donc compléter cette note d'avancement. Ce travail d'explicitation méthodologique doit favoriser la valorisation des données en les rendant transparentes, accessibles, interopérables et réutilisables. Une introduction générale précise le contexte scientifique dans lequel s'inscrit cette volonté de constitution de base de données standardisée sur le thème des événements de courses à travers le monde. La première partie présente les étapes de mise en œuvre de cette base de données.



https://running-datalab.com/

MOTS-CLES : DONNEES, CAPTATION, EXTRACTION, RUNNING, EVENEMENT, SCRAPING, CRAWLING, WEB, SOURCE NUMERIQUE, BASE DE DONNEES, COURSE A PIED.

[1] Programme de recherche dirigé par Mathilde Plard et financé pour la phase de structuration du jeu de données par le RFI TourismLab des Pays de la Loire.

CONTEXTE ET BESOINS

CONTEXTE INSTITUTIONNEL

Le programme de recherche vise à documenter l'actuel engouement pour les événements sportifs associés à la course à pied à l'échelle internationale. Le Running DataLab (RDL) s'occupe précisément de recenser et de qualifier ces événements. L'analyse ces événements et de leurs relations aux territoires permet d'explorer leur capacité à être force d'attractivité touristique et de développement territorial.

BESOIN

Pour analyser l'impact de ces événements de courses à pied sur les territoires, il est nécessaire de produire une base de données regroupant l'ensemble de l'offre des événements de courses à pied de façon la plus exhaustive. À partir de cette base de données, une cartographie descriptive est créée afin d'analyser la distribution spatiale des événements.

L'objectif de la manipulation est de créer une base de données sur les événements de courses à pied internationaux. Cette base de données doit : (1) être la plus exhaustive possible et (2) créer à partir de sources numériques hétérogènes.

EMPRISE SPATIALE ET TEMPORALITE

L'étude est réalisée à l'échelle internationale.

Il n'y a pas d'échelle temporelle définie : la collecte des événements de courses à pied peut concerner des événements passés ou à venir.

MÉTHODE

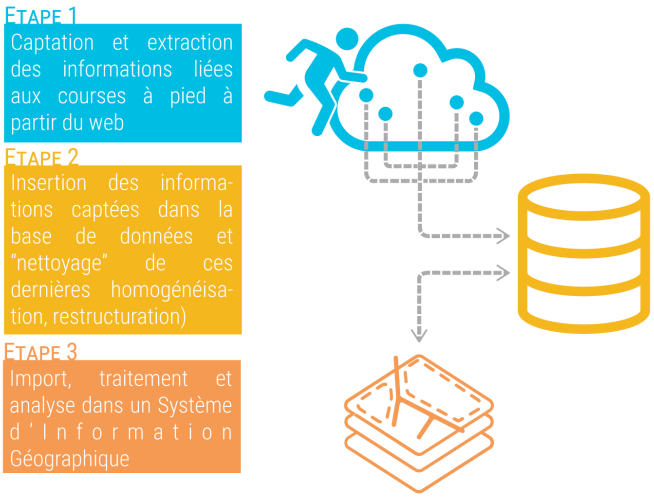
METHODOLOGIE GLOBALE

Pour alimenter la base de données, les données brutes sont captées puis extraites à partir de sources numériques hétérogènes. Il s'agit de sites internet qui proposent des agendas, soit une liste d'événement de course à pied. Ces sites captés et extraits sont nommés les sites sources.

Ici, le premier objectif est d'établir une méthodologie de captation/traitement des données à partir d'un site source. Il s'agit de tester une méthodologie qui soit interopérable à d'autres sites. Cette interopérabilité permettra de capter et d'extraire des données d'autres sites sources et de constituer une base de données plus complète d'un point de vue quantitatif et qualitatif.

Pour cette première étape de construction méthodologique, un site source est sélectionné. Les données textuelles sont ensuite captées et extraites. Importées dans un Système de Gestion de Base de Données Spatiale — SGBD (1), ces informations sont nettoyées, restructurées et traitées (2). Enfin, elles sont importées dans un Système d'Information Géographique (SIG) afin d'être spatialisées et analysées.

Méthodologie globale de la première étape



Conception/Réalisation : Violaine Guichet, Ingénieure d'études (2018)
Projet RunningDataLab, sous la direction de Mathilde Plard, chercheuse CNRS

DU WEB A LA BASE DE DONNEES

IDENTIFICATION DU SITE SOURCE

Rappel des objectifs

L'objectif final est de produire une cartographie descriptive des événements internationaux de courses à pied. Pour cela, il faut que les sites sources informent au minimum pour une course : du nom de l'événement, du nom de la course (ou du type de course) et de sa localisation (pays, état/région, ville).

Critères de sélection du site source

Trois critères ont permis de choisir le site source :

- quantitatif : nombre d'événements de courses à pied ;
- qualitatif : description de l'événement (nombre d'attributs), qualité des informations renseignées (précision, intérêt et redondance), emprise des données (internationale, régionale, nationale) ;
- technique :
    - Structuration HTML du site et faciliter d'extraction (absence de méthode AJAX, de scroll, etc.)
    - Structuration des informations détaillées (concaténation des dates, des localisations, etc.)

Site source retenu

Le site source retenu est celui de Marathon Ahotu. Ce dernier dispose d'un grand nombre de courses caractérisées par différentes informations (nom de l'événement, type de courses, tags, etc.). Son emprise est internationale.

LE SYSTEME DE CAPTATION-EXTRACTION DES DONNEES

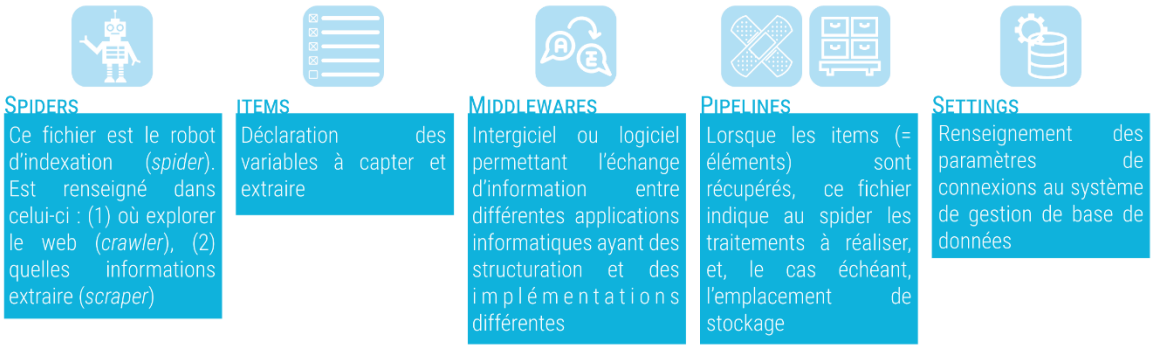
Outil de captation/extraction : la librairie Scrapy

Pour capter et extraire la donnée, la librairie Scrapy est retenue. Cette librairie, écrite en python, permet de « crawler » (naviguer de page en page) et de « scraper » (extraire des données).

Différents éléments sont à déclarer dans le script :

(1) les pages à crawler = les URL ;
(2) les informations à extraire = les balises HTML où sont situées les informations ;
(3) le lieu de dépôt des informations extraites = fichier csv ou SGBD
(4) la structuration des données à exporter = création des champs et typage (integer, string, etc.)

Structuration des fichiers de Scrapy



Rôle des fichiers Scrapy

Conception/Réalisation : Violaine Guichet, Ingénieure d'études (2018)
Projet RunningDataLab, sous la direction de Mathilde Plard, chercheuse CNRS

Définition des pages à crawler

Pour définir les pages à crawler, il fallait d'une part déclarer l'URL de base (https://marathons.ahotu.fr/calendrier) ainsi que les pages suivantes à crawler (https://marathons.ahotu.fr/calendrier/?page=2… https://marathons.ahotu.fr/calendrier/?page=5562).

Pour éviter d'avoir à renseigner les URL manuellement, une fonction en python a été mise en place pour décompter toutes les pages et ajouter au fur et à mesure les URL créées dans une liste. Cette liste contient alors toutes les URL à crawler.

Définition des balises à scraper

Pour extraire des informations, il faut spécifier les balises à scraper. Une page internet est structurée en bloc dans lesquels se situent des balises HTML qui contiennent différents types de contenu (texte, lien hypertexte, image, etc.). L'emplacement de la balise est indiqué par chemin Xpath.

Dépôt des données sur la base de données

Le script Scrapy dispose de différents fichiers ayant chacun une fonction.

Le lieu d'export est spécifié dans le fichier « pipelines.py ». Dans notre cas, l'export se fait dans un Système de Gestion de Base de Données (SGBD). Le logiciel utilisé est PostGreSQL 10, avec une cartouche spatiale postGIS 2.4.

Structuration des données exportées

Une première structuration des données captées est réalisée lors de l'intégration dans le SGBD.

Un SGBD structure les données en table qui sont composées d'attributs. Les tables sont en relation les unes avec les autres par un système de clés et d'identifiants uniques.

Le site exporté devient une table, chaque course devient une ligne soit une entité. 61 182 courses ont été exportées. Pour une course sont connus les attributs extraits suivants (=colonne) :

- Nom de l'événement
- Type de course (marathon, trail, 24 h, 5 K, etc.)
- Tag (course caritative, course féminine, montagne, désert, etc.)
- Localisation (concaténation de la ville, de la région et du pays)
- Date de l'événement
- URL de l'événement (lien vers la page dédiée à l'événement sur le site Marathon Ahotu, où il y a plus de détails)
- L'URL source (le site marathon Ahotu)

À ces attributs est ajouté un identifiant unique pour chaque course.

Description de la table des courses à pied exportées

Import_brut_Marathon_Ahotu
mar_id
mar_nom_evenement
mar_site_source
mar_URL_Even
mar_nom_course
mar_type_course
mar_tag

Ces informations sont spécifiques au site Marathon Ahotu. En fonction des sites internet listant les événements de type course à pied, les éléments varient :

- Informations listées pour un événement (nom événement, système de filtrage, tags, prix, nombre de participants, organisateur, etc.)
- Structuration des informations données : dans certains cas, la liste est faite en fonction de l'événement, pour d'autres, en fonction de la course, la localisation est nommée de différentes manières, le format de date change, etc.

- De la catégorisation des types de courses : il n'y a pas de catégorisation universelle des courses

TRAITEMENT DANS LA BASE DE DONNEES

ERREURS REMARQUEES SUR LES DONNEES IMPORTEES

Les données importées peuvent comporter différentes erreurs :

- Décalage dans les attributs,
- Ligne vide,
- Présence de retour chariot dans les cellules,
- Etc.

NETTOYAGE DES DONNEES IMPORTEES

Un nettoyage des données est opéré dans le SGBD :

- Vérification des attributs exportés,
- Suppression des lignes vides,
- Une requête SQL est lancée pour supprimer les doublons éventuels,
- Script pour supprimer les retours chariot.

GEOLOCALISATION DES EVENEMENTS

UTILISATION DE GEOPY

Le géocodage consiste à obtenir des coordonnées géographiques (longitude, latitude ou coordonnées x et y) à partir d'une adresse postale. Pour géocoder les courses, la librairie Geopy a été utilisée.

Geopy est une librairie en python simple d'utilisation puisque sa méthode (geocode) ne nécessite qu'un seul argument : l'adresse à géocoder. Geopy permet de mobiliser plusieurs services internet pour géocoder (Google Maps, OpenStreetMap, etc.). Dans notre cas, les courses ont été géocodées à l'aide d'OpenStreetMap grâce à la fonction Nominatim(). Il s'agit d'une implémentation de l'outil de recherche d'OSM. OpenStreetMap (OSM) est un projet collaboratif de création d'une base de données géographique à l'échelle mondiale. Toutes les données collectées sont réutilisables sous licence libre ODbL.

Cette librairie a été choisie, car contrairement à l'API Google Maps, il n'y a pas de limites dans le nombre de requêtes. Pour l'API Google Maps, le nombre de requêtes est limité à 2500 par jour.

PREPARATION DES DONNEES

Pour géocoder les courses, un fichier csv contenant leur liste est exporté du SGBD. Seules sont exportées deux colonnes : l'identifiant unique (ou id) et la localisation. Ce fichier est allégé des autres attributs.

Lors des premiers lancements du script, la présence de caractères spéciaux le fait boguer. La localisation est mise en majuscule pour supprimer ces caractères. La

localisation issue de l'extraction des courses est structurée de telle manière : « PAYS, REGION, VILLE » ou « PAYS, VILLE ».

Utilisation de Geopy

Pour faire fonctionner le script, différentes informations sont renseignées. Ces informations sont relatives, d'une part, aux adresses à géocoder, d'autre part, au fichier en sortie. Concernant les données en entrée sont précisés le chemin du fichier csv et l'emplacement de la colonne localisation au sein du fichier. S'agissant du fichier en sortie, doit être renseigné le nom du fichier, son format (csv) et sa composition (colonnes id, localisation, longitude et latitude).

Un autre argument a été modifié : le TimeOut. Cet argument spécifie le nombre de secondes à attendre si le service internet ne répond pas. Cette non-réponse peut être générée par différents facteurs (problème de limitation de requête, instabilité du réseau internet, etc.). Par défaut, l'argument est paramétré à 1 seconde, il a été changé à 60 secondes.

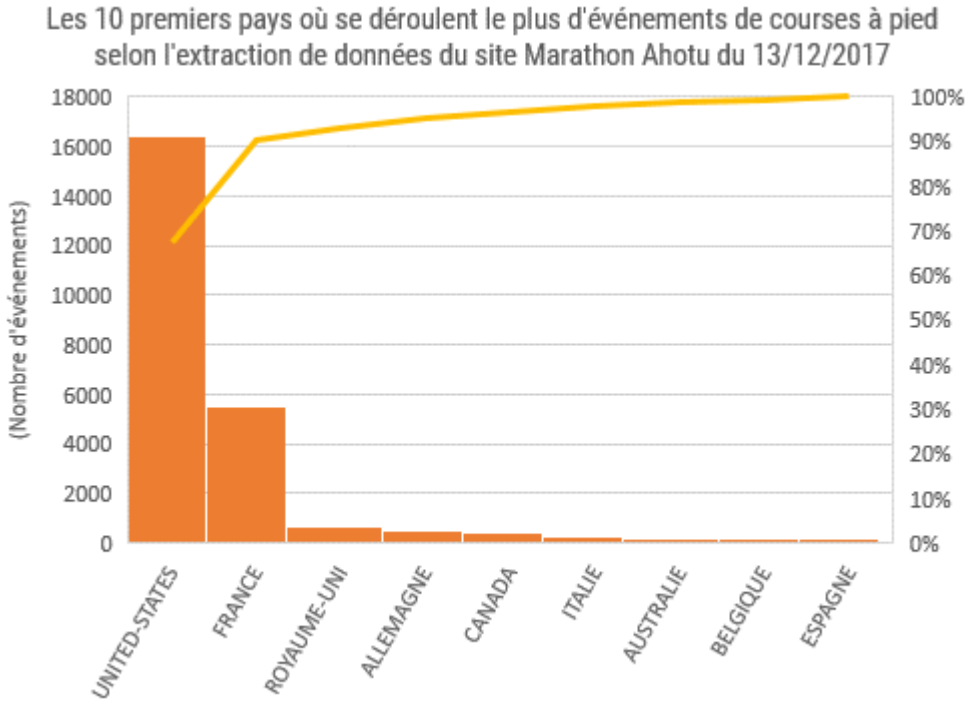
L'export a duré environ deux jours.

Resultats du geocodage

Sur les 61 182 courses extraites, 53 549 courses ont été géocodées soit 87,5 % de l'ensemble des courses. Ces courses sont divisées dans 26 176 événements. Ce qui donne une moyenne de course de 2,3 par événement. Les courses sont réparties sur 160 territoires.

Plus de 80 % des courses à pied extraites se localisent aux États-Unis (63 %) et en France (21 %).

Les événements se déroulent entre le 13 décembre 2017 (jour de l'extraction des données) et le 1er septembre 2019.



Source : site Marathon Ahotu

Réalisation : Violaine Guichet, Ingénieure d'études (2018)
Projet RunningDataLab, sous la direction de Mathilde Plard, chercheuse CNRS

MISE EN PLACE DE LA CARTOGRAPHIE DESCRIPTIVE DANS QGIS

GEOCODAGE ET STRUCTURATION DES DONNEES

Jointure des coordonnées x et y

Les courses géocodées sont importées dans la base de données. Une jointure est réalisée pour associer les coordonnées géographiques aux courses.

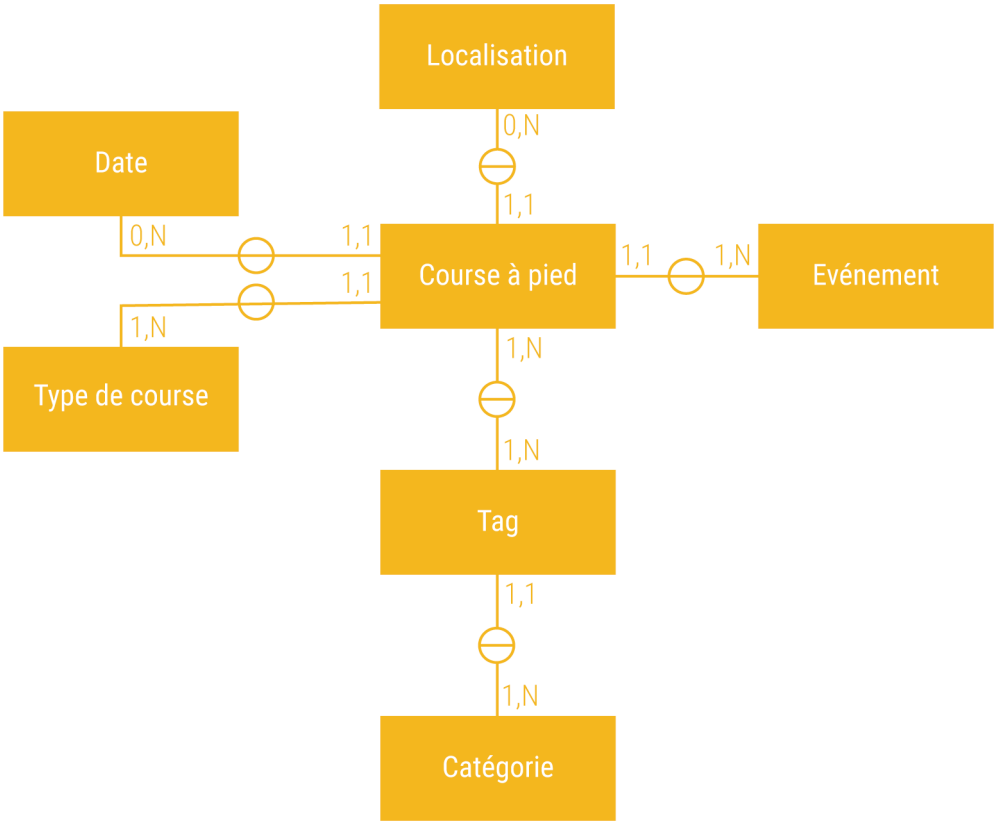
Modèle Conceptuel de Données simplifié du jeu de données

Une course géocodée possède un **type de course,** une **localisation** et une **date**. Les types de course sont les suivants :

- < 10 km.
- 10 km
- 10 km au semi-marathon
- 5 km
- Course à étape
- Course à X-heures
- Course sans-distance
- Course verticale
- Marathon
- Semi-marathon
- Semi-marathon au marathon
- Ultramarathon

Une course à pied peut être décrite par aucun ou plusieurs **tags** (mots-clés). Ces mots-clés sont très hétérogènes puisque le champ de saisie pour renseigner ces tags est libre. Ces mots-clés vont être catégorisés afin de créer des analyses thématiques. Un tag appartiendra à une catégorie.
Un **événement** de type running est composé d'une ou plusieurs courses à pied, en revanche, une course ne peut faire partir que d'un seul événement.

Modèle de données simplifié

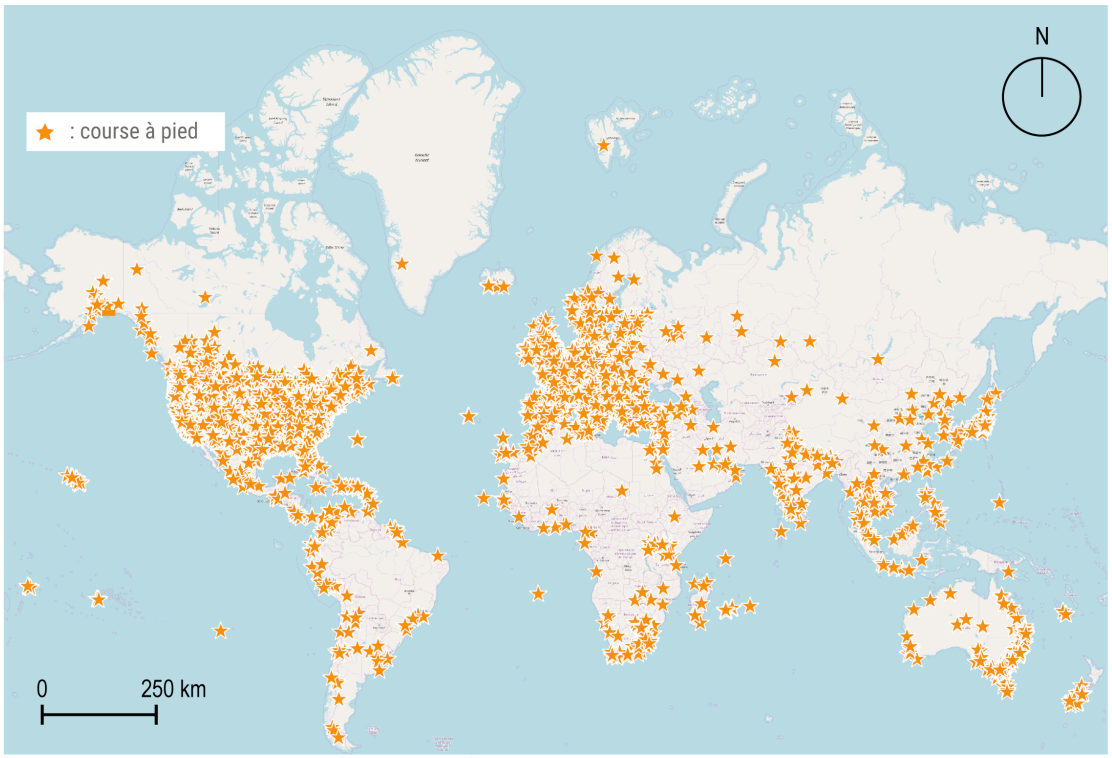


Réalisation/conception : Violaine Guichet, Ingénieure d'études (2018)
Projet RunningDataLab, sous la direction de Mathilde Plard, chercheuse CNRS

Connexion du SGBD avec un logiciel de Système d'Informations Géographiques (SIG)

Le SGBD est connecté au logiciel SIG (QGis 2.18.14) grâce à la cartouche spatiale postGIS de PostGreSQL qui permet d'intégrer des objets spatiaux dans la base de données. Lorsque les tables et les attributs sont importés dans QGis, une jointure attributaire est réalisée entre la table export_brut_marathon_ahotu contenant les courses extraites et le fichier csv de géocodage issu du script Geopy. La jointure attributaire est basée sur l'identifiant unique et permet d'intégrer les colonnes longitude et latitude à la table export_brut_marathon_ahotu.

Répartition spatiale des courses à pied extraites du web et géocodées



Sources : marathon Ahotu 13/12/2017, fond de plan : OSM,
Réalisation/conception : Violaine Guichet, Ingénieure d'études (2018)
Projet RunningDataLab, sous la direction de Mathilde Plard, chercheuse CNRS

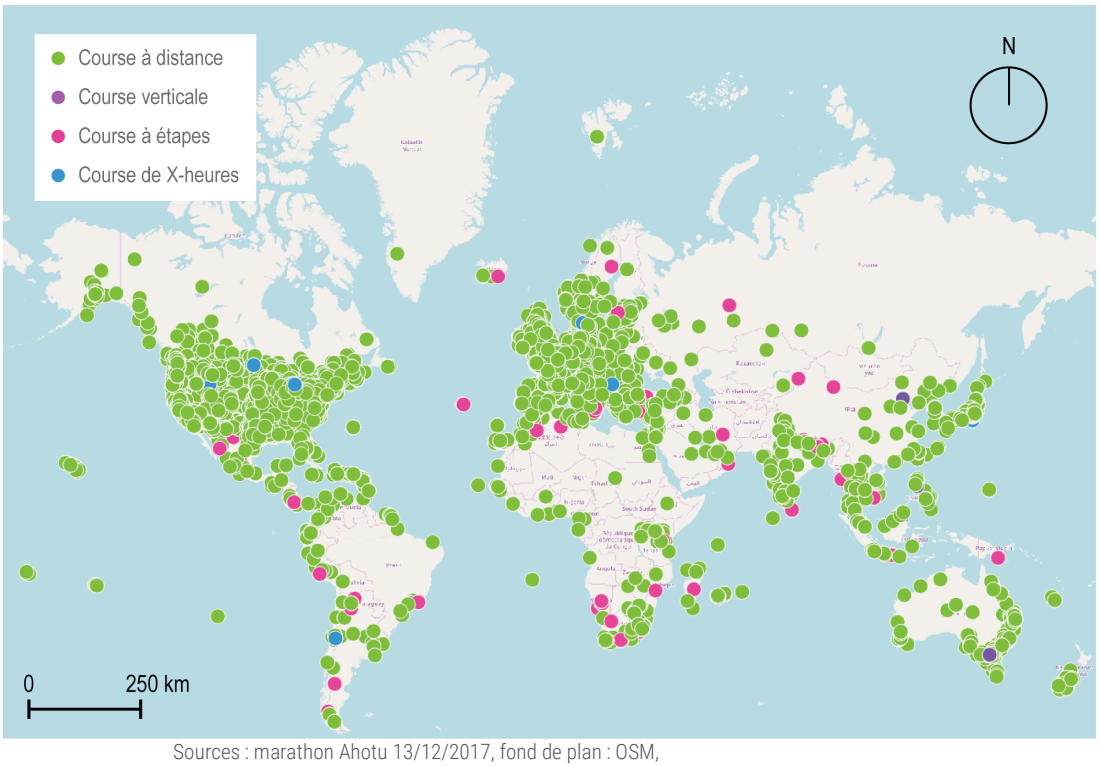
Realisation d'analyse thematique

Par type de courses

Dans le site, les différentes courses ne sont pas identifiées par des noms spécifiques, mais par leur paramètre. Ces paramètres sont au nombre de quatre :

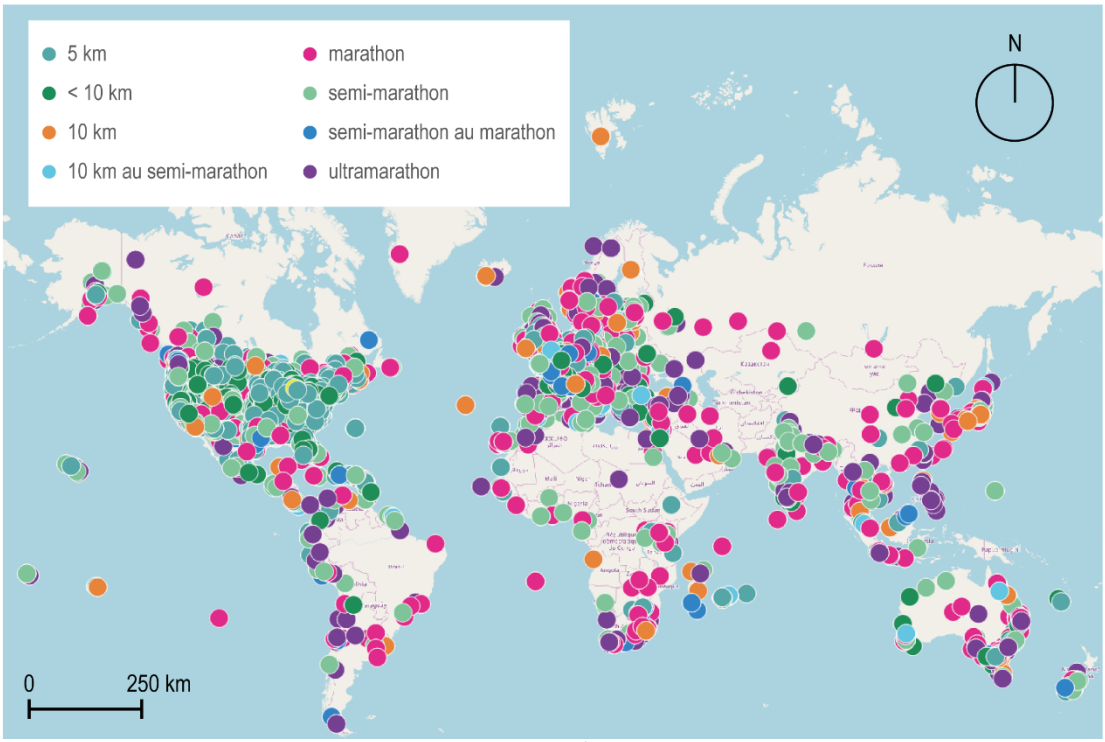
- Distance ;
- Dénivelé (= course verticale) ;
- Course à étape ;
- Temps (= course à X-heures)

Répartition spatiale des types de courses par paramètres



Sources : marathon Ahotu 13/12/2017, fond de plan : OSM,
Réalisation/conception : Violaine Guichet, Ingénieure d'études (2018)
Projet RunningDataLab, sous la direction de Mathilde Plard, chercheuse CNRS

Répartition spatiale des types de courses à distance

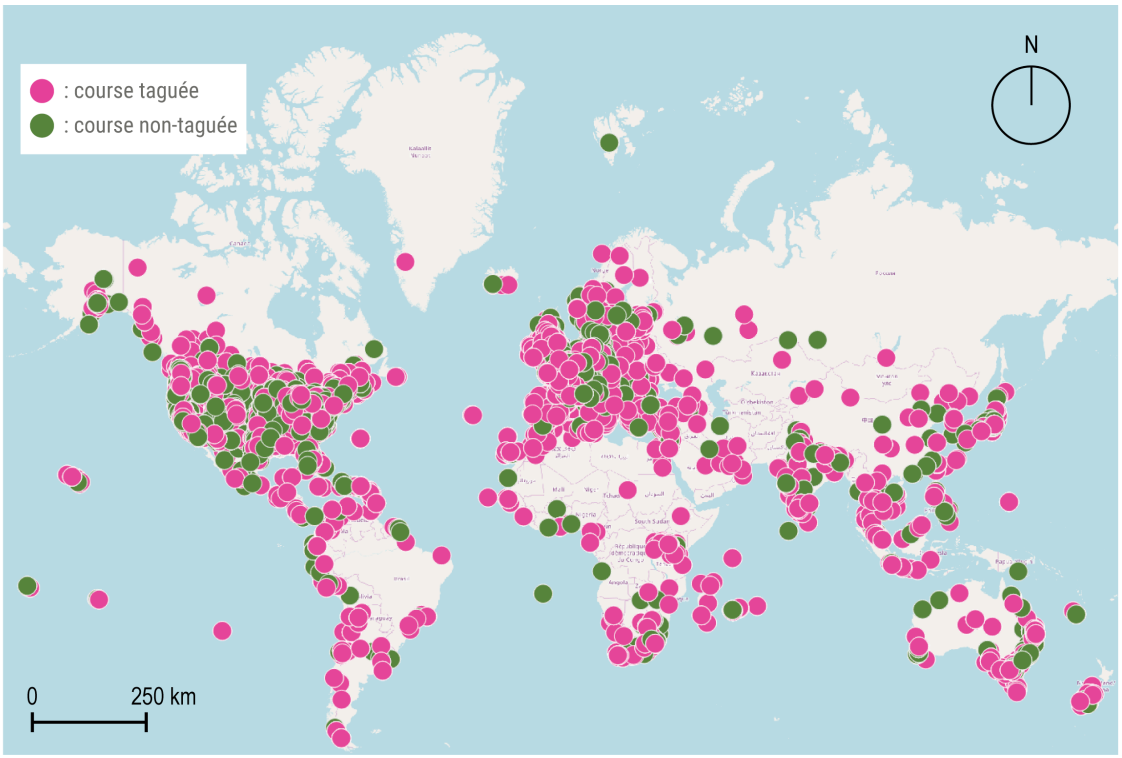


Sources : marathon Ahotu 13/12/2017, fond de plan : OSM,
Réalisation/conception : Violaine Guichet, Ingénieure d'études (2018)
Projet RunningDataLab, sous la direction de Mathilde Plard, chercheuse CNRS
Réalisation d'analyses thématiques à partir des « tags »

PLARD, GUICHET, 2018, « Captation, extraction et restructuration de données à partir de sources numériques hétérogènes » — Running DataLab

Pour décrire les courses à pied, le site internet utilise un système de « tags » (mots-clés). Sur les 53 549 courses à pied géocodées, 34 053 disposent de mots-clés.

Répartition spatiale des courses taguées ou non taguées



Sources : marathon Ahotu 13/12/2017, fond de plan : OSM,
Réalisation/conception : Violaine Guichet, Ingénieure d'études (2018)
Projet RunningDataLab, sous la direction de Mathilde Plard, chercheuse CNRS

Différentes analyses thématiques ont été réalisées pour catégoriser les courses à partir de ces tags. Cela a nécessité un travail de tri et de catégorisation.

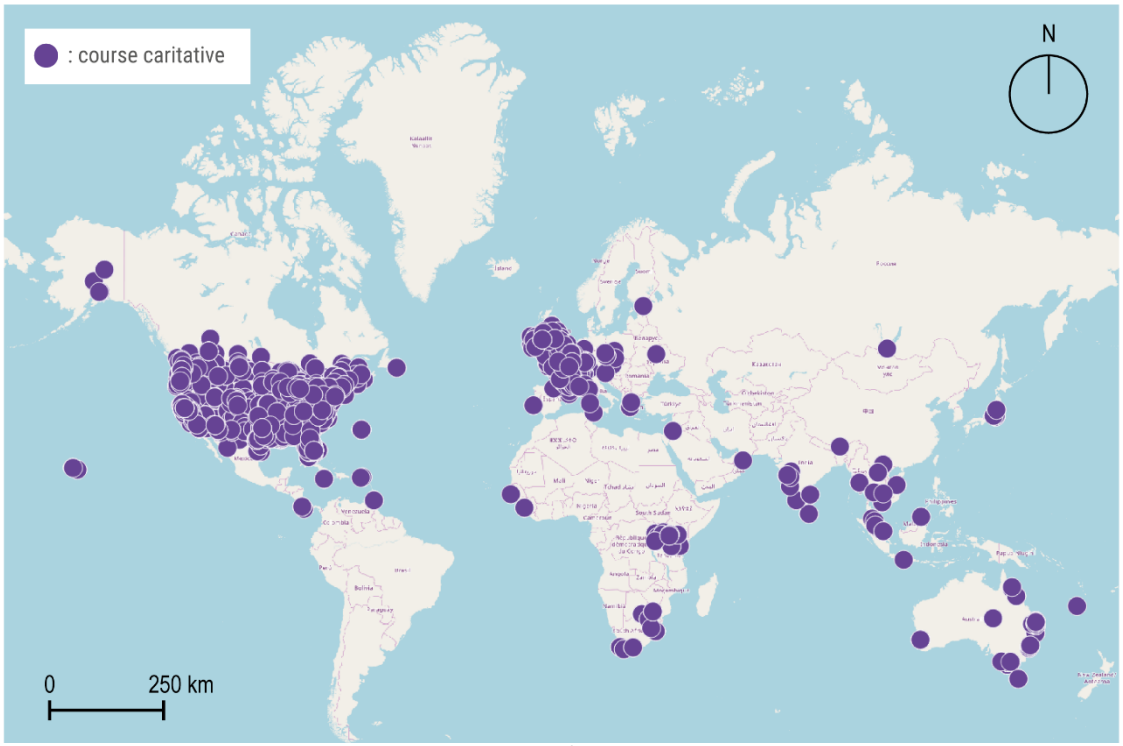
Le champ de saisi pour les tags est libre, cependant certains tags sont proposés par le site (ex. : « course caritative », « course féminine », etc.). Se trouvent alors deux classes de tags : ceux proposés par le site qui sont récurrents et dont l'écriture est homogène et une seconde classe de tags hétérogènes. Dans cette seconde classe de tags, la diversité est importante, les informations sont très hétérogènes allant de l'expression à des sigles et les langues diffèrent en fonction du pays d'accueil.

Un premier nettoyage dans les tags a été réalisé en ôtant :

- Les informations redondantes entre ce qui a déjà été renseigné pour une course et remis dans les tags (nom de l'événement, date)
- Les types de courses (ex. : 5 K, 10 K, marathon, ultra, etc.)
- Les mots génériques ont aussi été effacés (ex. : « The », « event », « Le », « run », etc.)

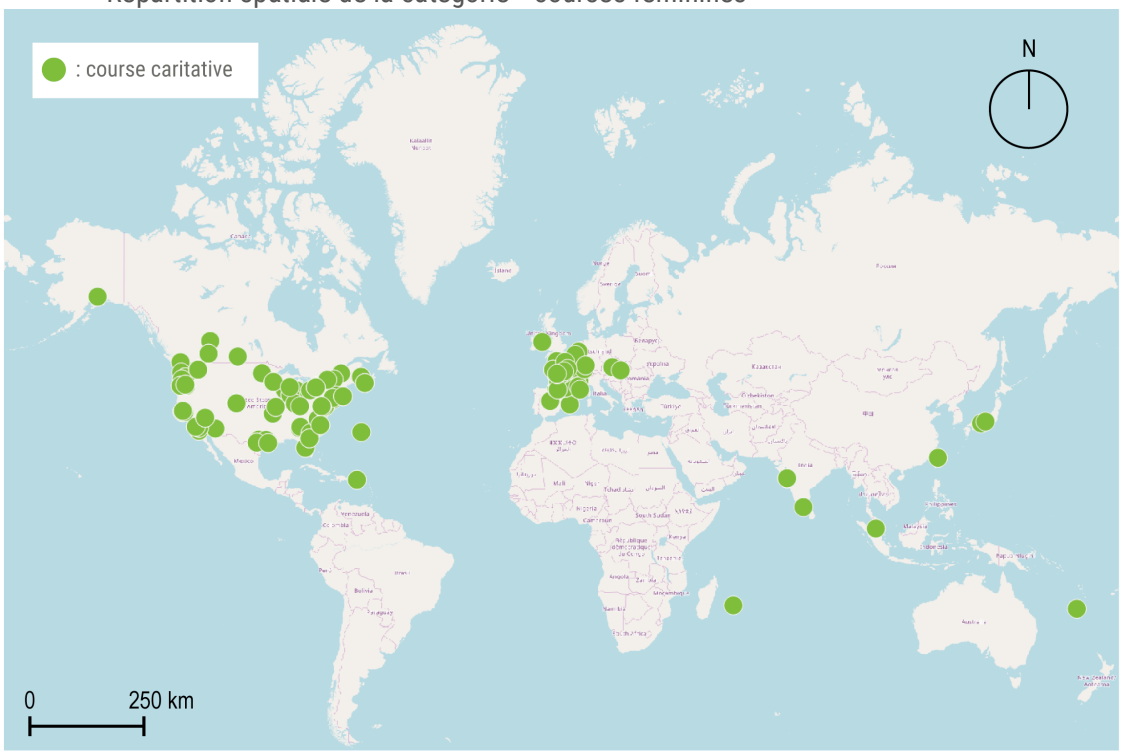
L'étape suivante est une étape de tri manuelle. Les mots-clés sont réunis dans différentes catégories pour créer des analyses thématiques. Par exemple, la catégorie course caritative regroupe les courses ayant des tags proposés par le site (« Bonne action », « Organisation caritative ») et d'autres tags en saisie libre (« Course à la vie », « A better World Running »). Un autre exemple, la catégorie féminine regroupe un tag proposé par le site (« course féminine ») et des tags libres (« DivaHalf », « Divas », « Girl », « Zooma »).

Répartition spatiale de la catégorie « courses caritatives »



Sources : marathon Ahotu 13/12/2017, fond de plan : OSM,
Réalisation/conception : Violaine Guichet, Ingénieure d'études (2018)
Projet RunningDataLab, sous la direction de Mathilde Plard, chercheuse CNRS

Répartition spatiale de la catégorie « courses féminines »



Sources : marathon Ahotu 13/12/2017, fond de plan : OSM,
Réalisation/conception : Violaine Guichet, Ingénieure d'études (2018)
Projet RunningDataLab, sous la direction de Mathilde Plard, chercheuse CNRS

PLARD, GUICHET, 2018, « Captation, extraction et restructuration de données à partir de sources numériques hétérogènes » — Running DataLab

LIMITES, AMELIORATIONS & PERSPECTIVES

LIMITE DE LA METHODOLOGIE DE CAPTATION/EXTRACTION

Un script à « usage unique »

Chaque site étant structuré différemment, le script écrit doit être réadapté pour chaque site (indication des balises, listing des différentes informations, etc.).

Reproduire cette méthodologie sur d'autres sites

L'objectif de ce travail est de fournir une cartographie la plus exhaustive possible de l'offre en running. En réalisant une vérification manuelle sur d'autres sites, a été constaté que l'offre de courses présente dans la base de données n'est pas exhaustive.

PERSPECTIVES & AMELIORATIONS

Insertion de nouveaux sites

Pour réaliser une base de données plus exhaustive sur la distribution spatiale des courses à pied, il apparait nécessaire de capter et extraire de l'information sur d'autres sites.

Actualisation et historisation des données

Pour alimenter et actualiser la base de données, cette opération de captation et extraction doit être reproduite pour acquérir les nouvelles courses à pied publiées sur le même site. Ce système permettra d'alimenter la base de données avec de nouvelles courses.

De plus, contrairement au site de type agenda d'événements de courses à pied, la base de données pourra archiver les données. Cette historisation permettra de constituer une base de données pour réaliser des analyses temporelles.

Automatiser plus les manipulations

Pour réduire le nombre de manipulations manuelles à réaliser, l'objectif est d'automatiser le plus possible de tâches, dans la mesure du possible. Dans cette logique, il serait possible d'insérer le code de géolocalisation dans un script SQL ou dans le script de captation/extraction Python dans du SQL. Ceci éviterait de devoir exporter la donnée, la réimporter et réaliser une jointure attributaire.

De même, le système d'actualisation pourrait être automatisé à l'aide d'un cron job.

DESCRIPTION DU JEU DE DONNEES

EMPRISE SPATIALE

Internationale

EMPRISE TEMPORELLE

De 2017 à 2019

DATE DE CREATION DU JEU DE DONNEES

Le jeu de données a été créé entre 2017 et 2018.

NOM DU FORMAT ET VERSION

Données géographiques : format shapefile

ORGANISATION DU JEU DE DONNEES

La couche cartographique contient les données suivantes :

	Nom	Type	
1	idu	String(254)	: identifiant unique
2	localisati	String(254)	: localisation (
3	latitude	Real(23,15)	: latitude
4	longitude	Real(23,15)	: longitude
5	URL_source	String(254)	: URL source [marathon-ahotu.fr]
6	typ_course	String(254)	: type de course [marathon, trail, 5 K, etc.]
7	nom_evenem	String(254)	: nom de l'événement
8	date	String(254)	: date où se déroule la course
9	jour	String(254)	: jour où se déroule la course
10	mois	String(254)	: mois où se déroule la course
11	annee	String(254)	: année où se déroule la course
12	URL_Even	String(254)	: URL de l'événement
13	tag	String(254)	: tags caractérisant les courses
14	MAJ	String(254)	: date de mise à jour de la donnée [13/12/2017]

CREATEUR DU JEU DE DONNEES

Violaine Guichet, Ingénieure d'études, projet Running DataLab sous la direction de Mathilde Plard.

COUT DE CREATION DES DONNEES

Temps travaillé [de la construction de la méthode aux cartographies] : équivalent à deux mois de travail en temps complet

Acquisition des données : gratuite

NOM DU JEU DE DONNEES

CAP_MarahonAhotu_Export1.shp