Construction de l'objet running à partir de sources numériques hétérogènes

Violaine Guichet, Mathilde Plard

CONSTRUCTION DE L'OBJET RUNNING 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 l'article doit favoriser la valorisation des données en les rendant accessibles, interopérables et réutilisables. Une introduction générale précise le contexte scientifique dans lequel s'inscrit cette volonté de structuration d'une 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.

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

https://running-datalab.com/

[1] Programme de recherche dirigé par Mathilde Plard (UMR 6590 CNRS) et financé pour la phase de structuration du jeu de données par le RFI TourismLab des Pays de la Loire.

CONTEXTE ET BESOIN

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 des événements de courses à pied sur les territoires, il est nécessaire de produire une base de données spatiale regroupant les événements de courses à pied internationaux. À partir de cette base de données spatiale, une cartographie descriptive est créée afin d'analyser la distribution spatiale des événements.

L'enjeu de ce projet est non seulement thématique (acquérir des données sur les événements de courses à pied pour analyser leurs impacts sociaux et territoriaux), mais aussi technique.

UN ENJEU TECHNIQUE

La base de données doit : (1) être la plus exhaustive possible et (2) être créée à partir de sources numériques hétérogènes. Il n'existe pas actuellement de base de données exhaustive sur les événements de courses à pied internationaux. Chaque structure (sportive, associative, etc.) dispose d'une masse d'informations relatives à sa thématique (ex. : FFA réunit des courses qualifiantes et/ou classantes). Les informations sur une course étant hétérogène d'une structure à l'autre.

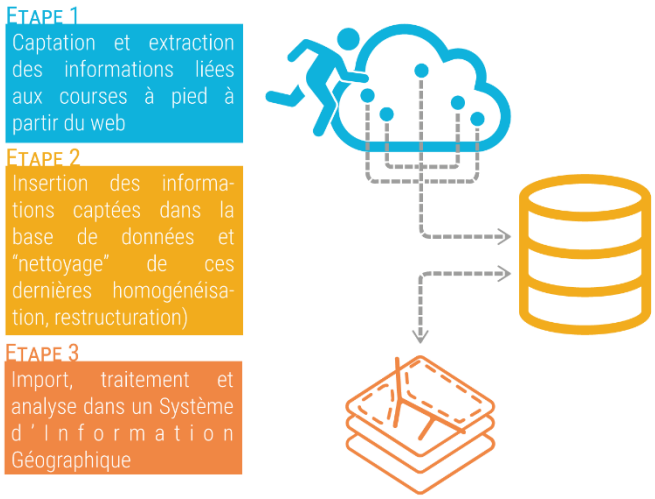
En cela, pour constituer la base de données, il faut recueillir ces données émanant de sources numériques hétérogènes pour un faire un jeu de données cohérent et exploitable.

METHODOLOGIE GLOBALE DE LA DEMARCHE

La base de données est constituée d'informations relatives aux événements de courses à pied internationaux.

La première étape consiste à extraire les informations de sources numériques hétérogènes : des sites sources. Dans un second temps, il faut nettoyer et homogénéiser le jeu de données extraites du web (étape 2). Enfin, un travail cartographique est réalisé dans un logiciel de Système d'Information Géographique pour réaliser des analyses thématiques en fonction des variables présentes dans le jeu de données (étape 3). Ce sont ces deux étapes qui sont décrites dans le présent papier.

Figure 1. Méthodologie de l'étape 1



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

MÉTHODE

DE SOURCES NUMERIQUES HETEROGENES A UN JEU DE DONNEES HOMOGENE

OBJECTIF

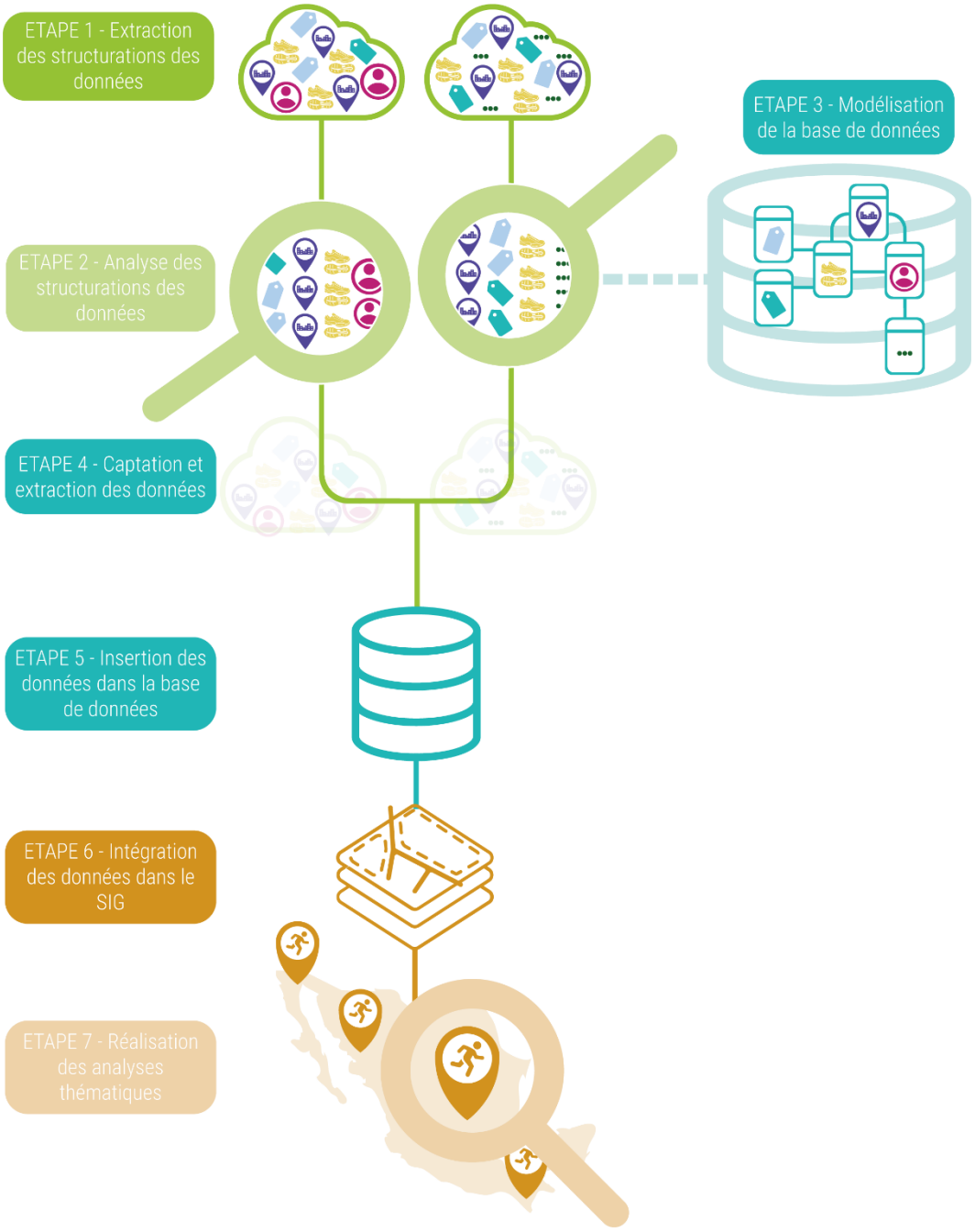
L'acquisition de données issues de plusieurs sites implique d'appréhender différents niveaux d'hétérogénéité concernant les jeux de données. D'une part, les données relatives aux courses diffèrent d'un site à l'autre (1). D'autre part, la manière dont sont structurées et déclinées les données varie (typologie des courses à pied, manière d'écrire la date, la localisation, etc.).

METHODOLOGIE GLOBALE

Cette méthodologie s'inscrit dans la continuité d'un travail engagé dans le projet du Running DataLab. La première étape de captation/extraction a été détaillée dans le papier intitulé « Captation, extraction et restructuration de données à partir de sources numériques hétérogènes » (V. Guichet, M. Plard), cette étape de captation/extraction ne sera donc pas décrite dans le présent papier.

Le nouvel enjeu décrit dans ce papier réside dans l'homogénéisation du jeu de données. Les données extraites émanent de différents sites sources. Ces sites disposent de sources hétérogènes : données diverses, règle de nommage variée, catégorisation différente, etc. L'objectif de restructurer les données extraites afin de produire un jeu de données cohérent permettant de faire des analyses thématiques sur l'ensemble des données.

Figure 1. **Description de la méthodologie globale**



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

RECUEIL DES STRUCTURATIONS DES DONNEES DANS LES SITES SOURCES

OBJECTIF

L'objectif est de restructurer des informations issues de sources hétérogènes pour en faire un jeu de données homogènes.

LES SITES SOURCES

Les données extraites sont issues de différents sites internet. Ils sont au nombre de cinq : jogging-plus, IAAF, Marathon Ahotu, Active, JustRunLah.

Le premier site est Marathon Ahotu, celui-ci avait été retenu lors de l'établissement de la méthodologie de captation/extraction, car il dispose d'un jeu de données important, d'une emprise internationale et d'une description des données structurée (système de catégorisation détaillée).

Ensuite, un deuxième site a été choisi pour des raisons plus techniques. Il s'agit du site active.com. Ce site est structuré selon les standards « Resource Description Framework » (RDF)[2]. Le modèle **RDF** permet de décrire de façon formelle les ressources du Web ainsi que leurs métadonnées (producteurs, date de création, etc.) : « [les] ressources [sont] identifiées par des URI[3], des propriétés et des classes (ou catégories) permettant de les définir, les décrire ou établir des relations entre elles »[4]. L'objectif était donc de faire un script type pour les données RDF qui soit réutilisable lorsque le site est structuré de cette manière. De plus, ce site contient aussi un jeu de données important, principalement tourné vers les pays anglophones.

Les trois autres sites ont été choisis en fonction de besoins thématiques. L'extraction des données à partir du site de l'IAAF permet de connaître le réseau de courses 'institutionnelles', soit les événements sportifs officiels internationaux. Les deux sites Jogging-plus et JustRunLah ont été retenus pour faire des études de cas plus précises sur certains territoires (Pays de la Loire et Asie).

CAPTATION ET EXTRACTION DES DONNEES

Deux méthodes d'extraction sont utilisées. La première méthode est celle semi-automatique qui fonctionne avec un script python (V. Guichet, M. Plard, 2018). Cette méthode est efficace lorsque le jeu de données est important et disparate au sein d'un site (plusieurs URL : onglets, pages, etc.). Son avantage réside dans le fait de pouvoir pointer les balises que l'on souhaite extraire. Aussi, cette méthode est pratique lorsque les structurations HTML sont complexes.

La seconde méthode, manuelle, consiste à venir simplement copier/coller les données d'un site source dans la base de données. Cette méthode fonctionne lorsque la structuration HTML du site est simple (structuration en tableau avec des balises de type <td>, <tr>) et que le jeu de données est concentré sur un nombre de pages limité (inférieur à 5).

---

[2] : https://www.w3.org/RDF/, site consulté le 28/03/2018

[3] : URI (Uniform Ressource Identifier) ou URL (Uniform Ressource Locator)

[4] : Documentation ABES - Agence Bibliographique de l'Enseignement Supérieur (http://documentation.abes.fr/aideidrefrdf/rdf.html), site consulté le 28/03/2018

Tableau 1. Méthode d'extraction selon les sites sources

LIBELLE_LONG	Méthode d'extraction	Nombre d'entités extraites
Jogging-plus	Manuelle	153
IAAF	Manuelle	105
Marathon Ahotu	Semi-automatique	53 534
Active	Semi-automatique	3 306
JustRunLah	Semi-automatique	547

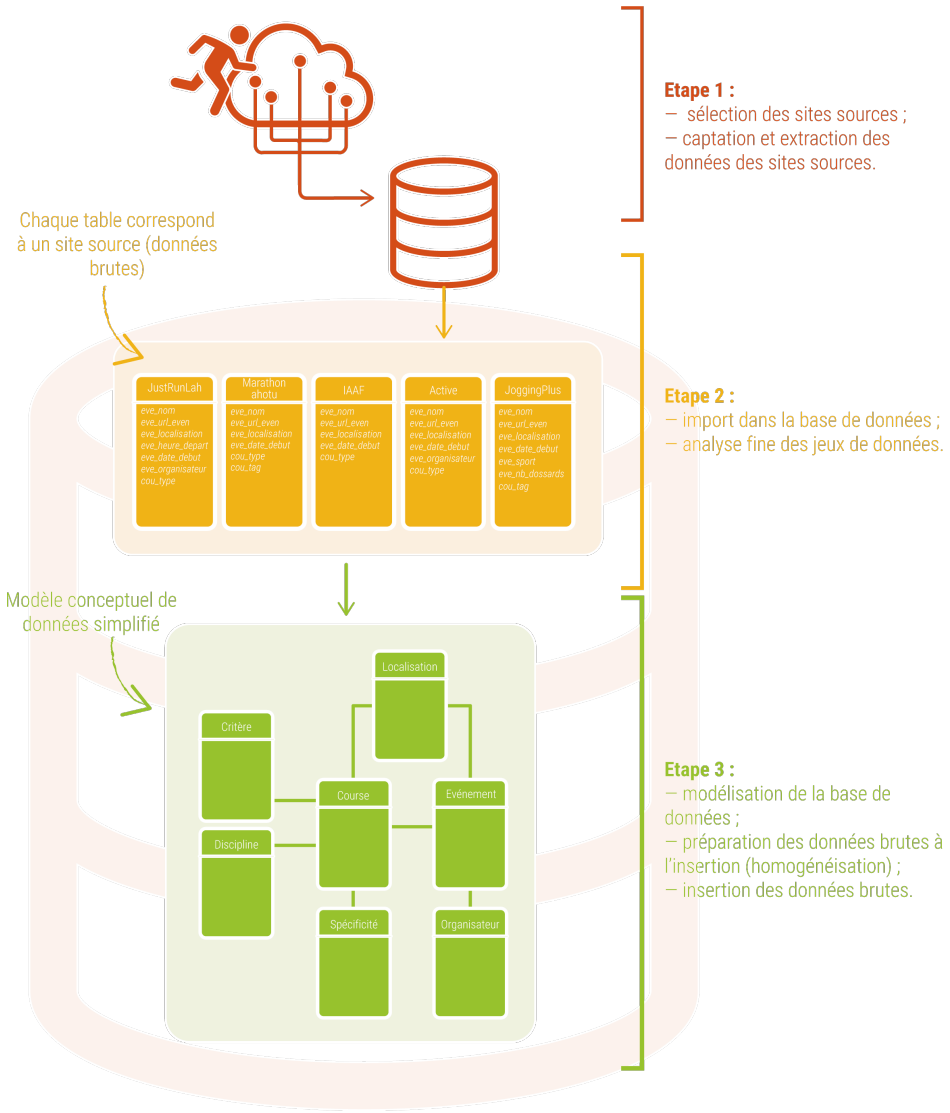
HOMOGENEISATION DU JEU DE DONNEES : LES ETAPES

Pour parvenir à homogénéiser le jeu de données, différentes étapes sont à prendre en considération, elles sont détaillées sur la figure ci-dessous (étape 2 et étape 3).

L'étape 2 est constituée de l'import des données dans le Système de Gestion de Base de Données (SGBD) et d'une analyse fine des jeux de données importées. Pour rappel, un SGBD structure les données en table qui sont composées d'attributs. Les tables sont en relation les unes avec les autres en fonction de leur lien. Cette relation est créée par un système de clés et d'identifiants uniques. Chaque site source exporté devient une table, chaque course (ou événement en fonction des sites) devient une ligne soit une entité.

L'analyse fine des jeux de données importées corrélées à une étude de la thématique (événements de courses à pied internationaux) permet de passer à l'étape 3 de conceptualisation de la base de données. Il s'agit de détailler les tables, leurs attributs et la connexion des tables les unes avec les autres. Une fois cette modélisation réalisée, la base de données est structurée et les données sont importées.

Figure 2. Méthodologie globale : détail sur les étapes d'homogénéisation



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

ÉTAPE 2 : ANALYSE DES JEUX DE DONNEES EXTRAITS

ÉTAT DES LIEUX SUR LES DONNEES DISPONIBLES EN FONCTION DES SITES SOURCES

En fonction des sites sources, les informations disponibles sont variées. A minima, pour les besoins de l'étude et ceux techniques, il faut connaître pour un événement : son nom, sa localisation, la date de début et les types de courses qui s'y déroulent. D'autres éléments sont présents dans les sites.

Tableau 2. Attributs des entités extraites selon les sites sources

Attributs	*Nom dans la table attributaire*	Jogging-plus	IAAF	Marathon Ahotu	Active	JustRunLah
Nom de l'événement	***eve_nom***	X	X	X	X	X
URL de l'événement	***eve_url_even***	X	X	X	X	X
Localisation de l'événement	***eve_localisation***	X	X	X	X	X
Date de début de l'événement	***eve_date_debut***	X	X	X	X	X
Type de sport de l'événement	***eve_sport***	X				
Nombre de dossards	***eve_nb_dossards***	X				
Heure de départ	***eve_heure_depart***					X
Organisateur	***eve_organisateur***				X	
Type de courses à pied	***cou_type***	X	X	X	X	X
Mot-clés, « tags » des courses à pied	***cou_tag***			X		

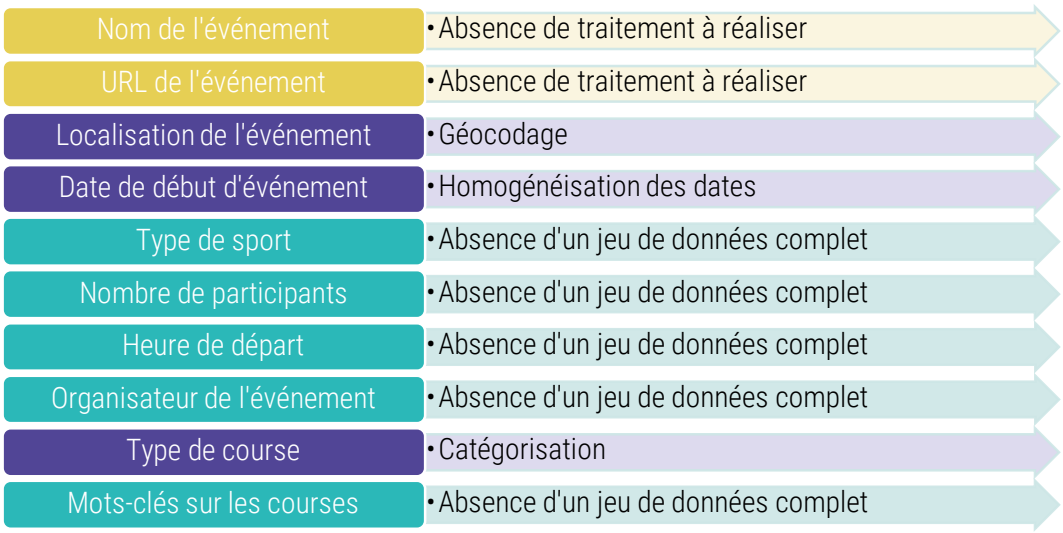
Les categorisations : absence d'homogeneite entre les sites sources

Concernant les éléments eve_sport, cou_type et cou_tag, les systèmes de sont variés : structuration spécifique à des sites, champs de saisie libre, langue, etc. (cf. Tableau 1 en annexe : « Liste des attributs selon les sites sources »). Il apparait nécessaire de recréer des catégorisations qui prennent en compte toutes ces catégories.

Points sur les attributs exploitables

Au total, il existe neuf attributs pour les événements et courses. Ces attributs ne sont pas renseignés dans tous les sites. Pour constituer un jeu de données homogènes, ne pourront être mobilisés que les attributs ayant été renseignés dans tous les sites (cf. Illustration ci-dessous éléments en violet). D'autres attributs ne nécessitent pas de traitement (cf. Illustration ci-dessous éléments en jaune).

Figure 1. Traitement des attributs



Concernant les attributs mobilisables, deux types de traitements vont être réalisés :

- Un travail d'homogénéisation : sur (1) la date et (2) le type de course ;
- Un travail de localisation : géocodage des événements ou courses à pied.

ÉTAPE 3 : STRUCTURATION DU NOUVEAU JEU DE DONNEES

MODELISATION DE LA BASE DE DONNEES

Les catégorisations de courses à pied

De la même manière que pour la base de données, les catégorisations ont été réalisées à partir d'une approche empirique et théorique.

Déterminer et qualifier les sports de courses à pied

Les courses à pied sont une discipline rattachée à l'athlétisme (source : Fédération Française d'Athlétisme – FAA). Cette discipline dispose de différents sports ayant eux-mêmes leurs critères qui les permet de les différencier.

Ce travail a été réalisé à partir de trois sources : la Fédération Française de TRIathlon (FFTRI), la Fédération Française d'Athlétisme (FFA), et les données extraites des 17 sites sources. Cela a permis de :

- Lister exhaustivement tous les sports de courses à pied
- Définir leur critère

En fonction des courses, les critères sont plus ou moins précis. Cette précision diffère selon la labellisation des courses. Une course labellisée par la FFA ou la FFTRI nécessite d'être bien définie pour que les critères soient respectés et que d'une course à l'autre, les modalités soient les mêmes.

Ouverture de la catégorisation aux courses à pied non officielles

En revanche, pour des courses non labellisées, les critères peuvent évoluer d'un événement à l'autre. Cependant, du fait de la fréquence et de la popularité de certaines courses à pied, certains critères viennent à être officialisés. C'est le cas des color run qui, malgré une non-labellisation, se définissent par une distance de 5 km et des lancers de pigments de couleurs tous les kilomètres.

Des critères prépondérants

L'élément qualificatif le plus récurrent pour répartir les courses dans les sites est la distance. Cet élément est le point d'entrée dans la catégorisation. Des épreuves ou disciplines sont distinguées explicitement dans tous les sites (triathlon, duathlon, course X-heures, verticale, etc.). De ce fait, ces courses peuvent faire l'objet d'une catégorisation spécifique.

Présentation de la structuration de la base de données

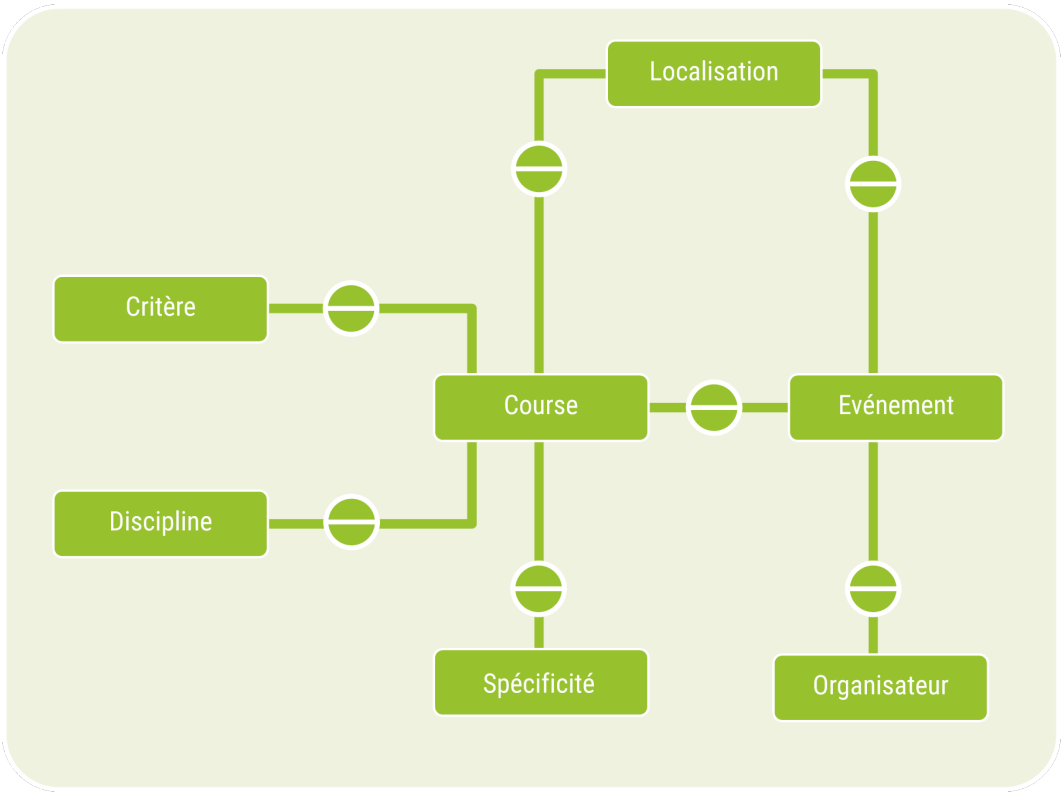
La base de données est modélisée à partir de deux approches : l'une empirique émanant des résultats des sites sources, une autre plus théorique qui résulte d'un travail de définition des termes. Les tableaux n°1 et 2 et la figure n°1 en annexes illustrent ce travail de définition.

Ci-dessous en figure n°5 est présentée la structuration de la base de données. Elle permet d'organiser les éléments inclus dans le projet de recherche et de détailler les éléments les uns avec les autres.

Ici, un événement de courses à pied peut contenir une ou plusieurs courses à pied. Il est planifié par un ou des organisateurs et est localisé par un lieu.

La course à pied quant à elle est rattachée à un événement. Elle peut avoir une localisation différente de l'événement, puisque les points de départ de courses peuvent varier de celui d'implantation de l'événement. La course est définie par un certain nombre de critères (distance, support – sentier, route, etc., dénivelé, équipe, temps). De même, une course peut être rattachée à une discipline et sous-discipline (niveau 1 et 2 du tableau n°2 situé en annexe) et peut avoir des spécificités (courses féminine, nocturne, extrême, etc.).

Figure 3. **Modèle Conceptuel de Données simplifié**



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

IMPORT DES DONNEES BRUTES DANS LES TABLES

TABLES « EVENEMENT » ET « COURSE »

Les données sont insérées dans deux types de tables. Une table « evenement » et une table « course ». En fonction des sites, les données sont listées soit par course, soit par événement. Durant un événement peuvent se dérouler différentes courses.

Seul le site de Marathon Ahotu présente son jeu de données par course. Pour ce site, les données sont insérées directement dans la table « course ». Chacune des courses est présentée par le nom de l'événement. Pour avoir la liste des événements, les doublons sont retirés en fonction du nom de l'événement ce qui permet d'avoir la liste unique des événements et d'insérer les données dans la table « evenement ».

Les autres sites sont listés par événements. Les types de courses sont décrits, mais concaténés dans une même cellule. Les données sont dans un premier temps insérées dans la table « evenement ». Pour insérer les courses dans la table course, les données sont déconcaténées, afin d'avoir pour une course par ligne.

Tableau 1. Passage d'une structuration en événement en course

1. Avant donnés dans la table « evenement »

Nom événement	Type de course
Royal Canal Run Longford 2018	10K, ULTRA, MARATHON, HALF MARATHON

2. Création de la table événement et déconcaténation

Nom événement	Type de course
Royal Canal Run Longford 2018	10K
Royal Canal Run Longford 2018	ULTRA
Royal Canal Run Longford 2018	MARATHON
Royal Canal Run Longford 2018	HALF MARATHON

L'ATTRIBUT DATE

Pour réaliser cette homogénéisation, les éléments des dates ont été décomposés pour être répartis dans trois colonnes (jour, mois et année). Dans les colonnes jour et mois, les éléments textuels ont été traités pour devenir des éléments numériques. Aucun traitement n'a été réalisé sur la colonne année. In fine, la date est structurée de telle manière : JJ/MM/AAAA.

GEOCODAGE

Méthode

Pour géocoder, la librairie utilisée est Geopy. Cette librairie permet de géocoder différentes adresses en mobilisant des APIs variées (Google ou OpenStreetMap par exemple). L'objectif est de trouver à partir des localisations détenues pour chaque entité les coordonnées géographiques afin de réaliser les cartographies.

Pour le géocodage des données extraites du site Marathon Ahotu, la fonction Nominatim() avait été utilisée. Cette fonction fait appel à l'API d'OpenStreetMap (OSM). Ce choix avait été réalisé, car, contrairement à l'API Google qui limite le nombre de requêtes à 2500 par jour, l'API d'OSM n'a pas de limitation.

En revanche, la totalité des entités extraites des trois sites restants étant inférieure à 2 500, les deux fonctions ont été utilisées pour pouvoir comparer les résultats.

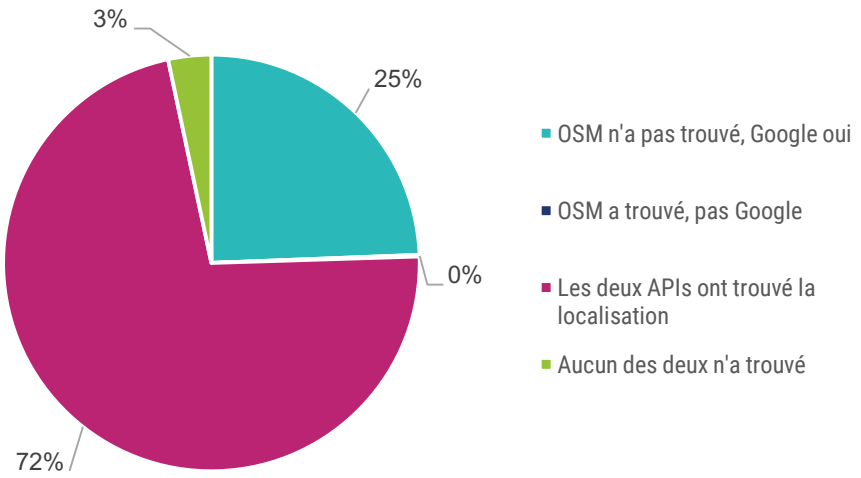
Tableau 1. API utilisée pour le géocodage selon les sites sources

	API utilisée pour le géocodage	
	OSM API	Google API
Jogging-plus	X	X
IAAF	X	X
Marathon Ahotu	X	
Active	X	
JustRunLah	X	X

Comparaison des géocodages

Au total, les entités à géocoder sont au nombre de 808. Concernant l'API OSM, 584 entités ont été géocodées ce qui représente un taux de géocodage de 72,3%. Avec l'API Google, 96,5% des entités ont été géocodées soit 780 entités. Dans les trois quarts des cas, les deux APIs ont géocodé l'entité. Une fois sur quatre, l'API Google a géocodé l'entité et pas l'API d'OSM. Il n'y a qu'un cas de figure où l'API d'OSM a trouvé les coordonnées géographiques et non l'API de Google.

Figure 1. Comparaison du géocodage entre l'API de Google et celle d'Open Street Map

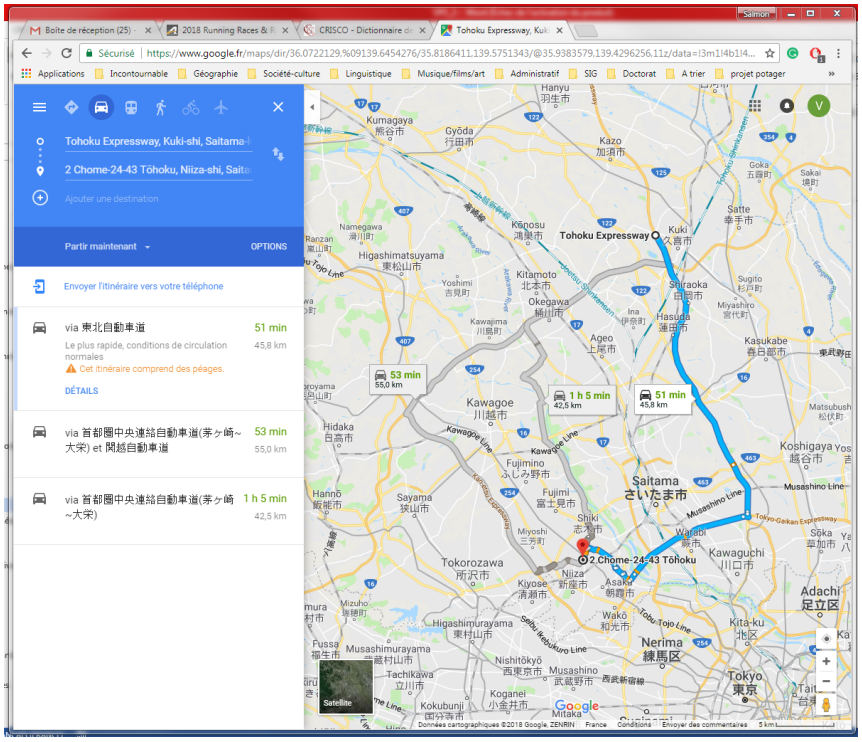


Conception/Réalisation : Violaine Guichet, Ingénieure d'études (2018)
Projet Running DataLab, sous la direction de Mathilde Plard, chercheuse CNRS

Lorsque les deux APIs ont trouvé la localisation, les coordonnées diffèrent. Cette différence peut être de quelques mètres ou kilomètres.

Figure 2. Exemple de géocodage de l'événement 351_JUSTRUNLAH

Point rouge Google / autre point OSM : 45 km de différence.



Sources : projet Running DataLab et GoogleMap

Cette différence dans les coordonnées géographiques relève d'une part du système de géocodage de chacune des APIs. Par exemple, pour des cas où seuls les pays sont donnés, les coordonnées diffèrent. D'autre part, la différence de géocodage relève de la qualité de la localisation de départ (texte dans la table attributaire). Plus la localisation est précise, plus le géocodage l'est aussi. Pour évaluer ce biais, une notation de la précision de l'adresse de départ a été mise en place dans la table attributaire ce qui permet lors de la mobilisation des données de prendre en compte cet élément :

Tableau 1. Notation de la précision géographique des entités extraites

	Précision géographique	
	Note	Description
Très bonne	5	Adresse précise (nom de voie, ville, région/état, pays)
Bonne	4	Localisation à la ville (ville, région/état, pays)
Moyen	3	Localisation à la région/état (région/état, pays)
Médiocre	2	Localisation au pays (pays, région du globe)
Très médiocre	1	Localisation au continent (région du globe)
Nul	0	Absence de localisation

En fonction des études où le jeu de données est mobilisé, cette codification peut être mobilisée. Par exemple, dans le cadre d'une étude de cas à l'échelle régionale, il est conseillé d'utiliser seulement les entités dont la note est égale ou inférieure à 3. En revanche, dans le cadre d'une étude internationale sur la répartition spatiale des courses à pied, il peut être utilisé les notes de 5 à 2 ou 1.

GESTION DES DOUBLONS

CONTROLE SELON LE NOM DE L'EVENEMENT

Le premier contrôle pour vérifier s'il existe des événements en doublons est réalisé sur le nom de l'événement et la date de début.

Cette vérification ne suffit pas pour contrôler l'entièreté des événements. D'un site à l'autre, les noms peuvent être modifiées ou des éléments ajoutés (année de déroulement de la course). Un deuxième contrôle est réalisé.

CONTROLE SELON LA LOCALISATION ET LA DATE DE DEBUT D'EVENEMENT

Un contrôle sur la localisation et la date de début d'événement est réalisé. La localisation prise n'est pas celle textuelle, mais celle issue du géocodage, soit les coordonnées x et y. Il a été vu précédemment qu'en fonction d'un géocodage avec l'API de Google ou celui d'OSM, les coordonnées pouvaient être différentes. Pour réaliser la vérification, seules les coordonnées x et y issues de l'API d'OSM sont utilisées puisqu'il s'agit de celles présentes sur l'ensemble du jeu de données.

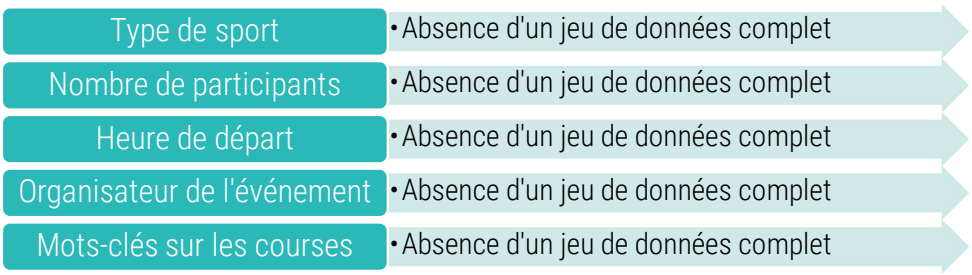
LIMITES, AMELIORATIONS & PERSPECTIVES

CONTROLE DE L'EXHAUSTIVITE QUANTITATIVE DU JEU DE DONNEES

Comme explicité précédemment, il n'existe pas de base de données exhaustive concernant les événements de courses à pied internationaux. En ce sens, il n'est pas possible d'évaluer l'exhaustivité quantitative de la base de données créée. Pour l'évaluer, il faudrait réaliser un inventaire exhaustif à partir d'une étude de cas locale, en mobilisant différents acteurs territoriaux par exemple.

AMELIORER LA QUALIFICATION DES EVENEMENTS

Toutes les informations ne sont pas détaillées sur l'ensemble des sites sources. Il faudrait continuer ce travail de qualification sur les attributs non homogènes :



Mettre en place un systeme d'actualisation automatique pour les methodes semi-automatiques

Afin d'alimenter la base de données, il serait intéressant de mettre en place un outil d'actualisation automatique.

Évaluation du nombre de doublons encore present

Malgré les deux vérifications sur les doublons, il est possible que des doublons subsistent. En effet, la précision de la localisation textuelle diffère d'un site à l'autre. Certains indiquent les bâtiments, alors que d'autres ne donnent que la ville voire l'état. Afin d'avoir une vérification plus poussée, il faudrait vérifier manuellement les données et évaluer la part de doublons potentiels sur l'ensemble du jeu de données.

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

Createur du jeu de donnees

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

Cout de creation des donnees

Temps travaillé : équivalent de trois semaines en temps plein

Nom du jeu de donnees

CAP_MarahonAhotu_Export1.shp

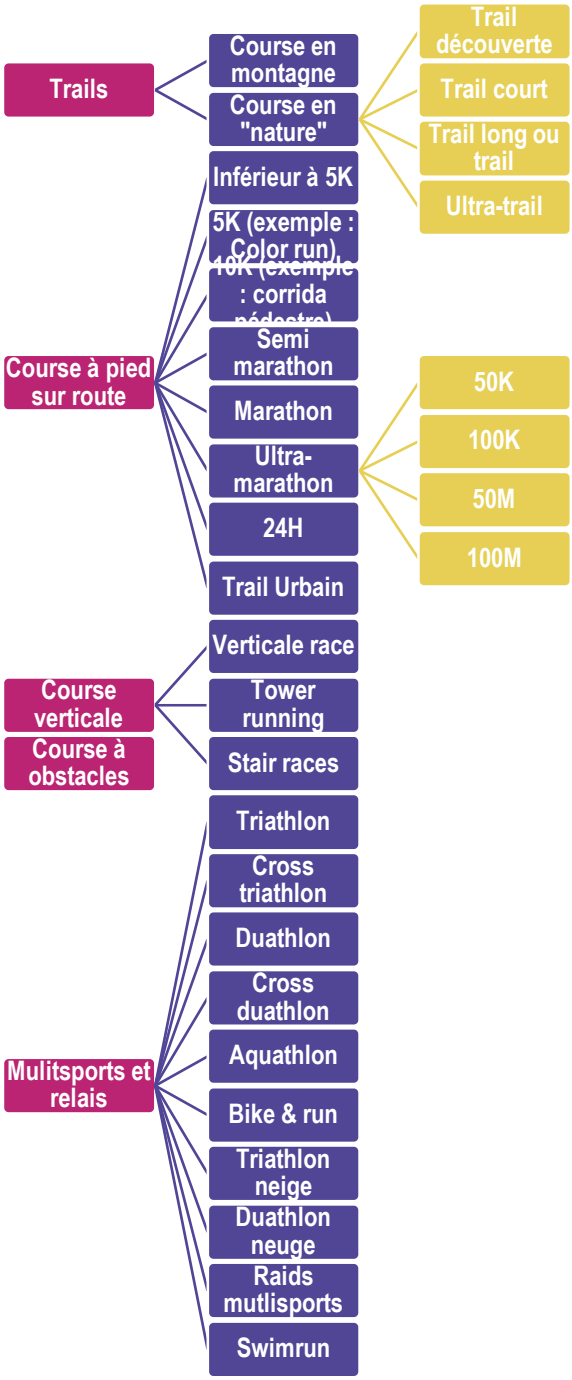
Type de donnees

Langues

Français et anglais principalement.

ANNEXES

Tableau 1. **Catégorisation des sports de courses à pied**



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

Tableau 2. Déclinaison des épreuves de courses à pied, des critères et des sources

Catégorisation			Critères de définition						Source de la définition
Niveau 1	Niveau 2	Niveau 3	Distance 1	Distance 2	Distance 3	Dénivelé (+, en m)	Temps min.	Temps max.	
Les trails	Course en montagne					500	1h	1h15	FFA
	Course en "nature"	Trail découverte	0 à 20 km						FFA
		Trail court	21 à 42 km						FFA
		Trail long ou trail	42 à 80 km			2000			FFA
		Ultra-trail	plus de 80 km			plus de 2000			FFA
Course à pied sur route	Inférieur à 5K		inf à 5 km						FFA
	5K	Divers	5 km						FFA
		Color run							Analyse des différents événements
	10K	Divers	10 km						
		Corrida pédestre							Analyse des différents événements
	Half marathons		21,097 km						FFA
	Marathons		42,195 km						FFA
	Ultra Marathons	50K	50 km						FFA
		100K	100 km						FFA
		50M	50 miles						FFA
		100M	100 miles						FFA
	24 heures		1km						FFA
	Trail urbain								Analyse des différents événements
Course verticale	Verticale race					1000			Analyse des différents événements
	Tower Running								Analyse des différents événements
	Stair races								Analyse des différents événements

Course à obstacles	Les courses à obstacle classiques								Analyse des différents événements
	Courses à obstacles spécifiques	Tough Muder							Analyse des différents événements
Multisport et relais	Triathlon (natation/cyclisme/CAP)	XS	400m	10km	2.5km				FFTRI
		S	750m	20km	5km				FFTRI
		M	1.5km	40km	10km				FFTRI
		L	3km	80km	20km				FFTRI
		XL	4km	120km	30km				FFTRI
		XXL	3.8km	180km	42.195km				FFTRI
	Cross triathlon (natation/cyclisme/CAP)	XS	250m	5.5km	2km				FFTRI
		S	500m	11km	4km				FFTRI
		M	1km	22km	8km				FFTRI
		L	2km	44km	16km				FFTRI
		XL	3km	66km	24km				FFTRI
	Duathlon (CAP/cyclisme/CAP)	XS	2.5km	10km	1.25km				FFTRI
		S	5km	20km	2.5km				FFTRI
		M	10km	40km	5km				FFTRI
		L	10km	80km	10km				FFTRI
		XL	20km	120km	10km				FFTRI
		XXL	20km	180km	20km				FFTRI
	Cross duathlon (CAP/cyclisme/CAP)	XS	2km	5.5km	1km				FFTRI
		S	4km	11km	2km				FFTRI
		M	8km	22km	4km				FFTRI
		L	16km	44km	8km				FFTRI
		XL	24km	66km	12km				FFTRI
	Aquathlon (natation/CAP)	XS	500 m	2.5km					FFTRI
		S	1km	5km					FFTRI
		M	2km	10km					FFTRI
		L	3km	15km					FFTRI
		XL	4km	20km					FFTRI
	Bike & Run	S					0h45 et 1h15		FFTRI
		M					1h15 et 2h		FFTRI
		L					plus de 2h		FFTRI
	Triathlon neige (natation/cyclisme/CAP)	XS	2 km	3.5 km	3 km				FFTRI
		S	4 km	7 km	6 km				FFTRI
		M	8 km	14 km	12 km				FFTRI
	Duathlon neige (natation/cyclisme/CAP)	XS	2 km	3 km	2 km				FFTRI
		S	4 km	6 km	4 km				FFTRI
		M	8 km	12 km	8 km				FFTRI

	Raids multisports	XS					moins de 5h		FFTRI
		S					entre 5 et 7h		FFTRI
		M					entre 7 et 12h		FFTRI
		L					entre 12 et 24h		FFTRI
		XL					entre 24 et 36h		FFTRI
	Swimrun	XS	Inf à 5 km						FFTRI
		S	entre 5 et 12,5 km						FFTRI
		M	entre 12,5 et 20 km						FFTRI
		L	entre 20 et 35 km						FFTRI
		XL	entre 35 et 55 km						FFTRI
		XXL	plus de 55 km						FFTRI

Source : Projet Running DataLab — Guichet, Plard, 2018

Tableau 3. Liste des attributs selon les sites sources

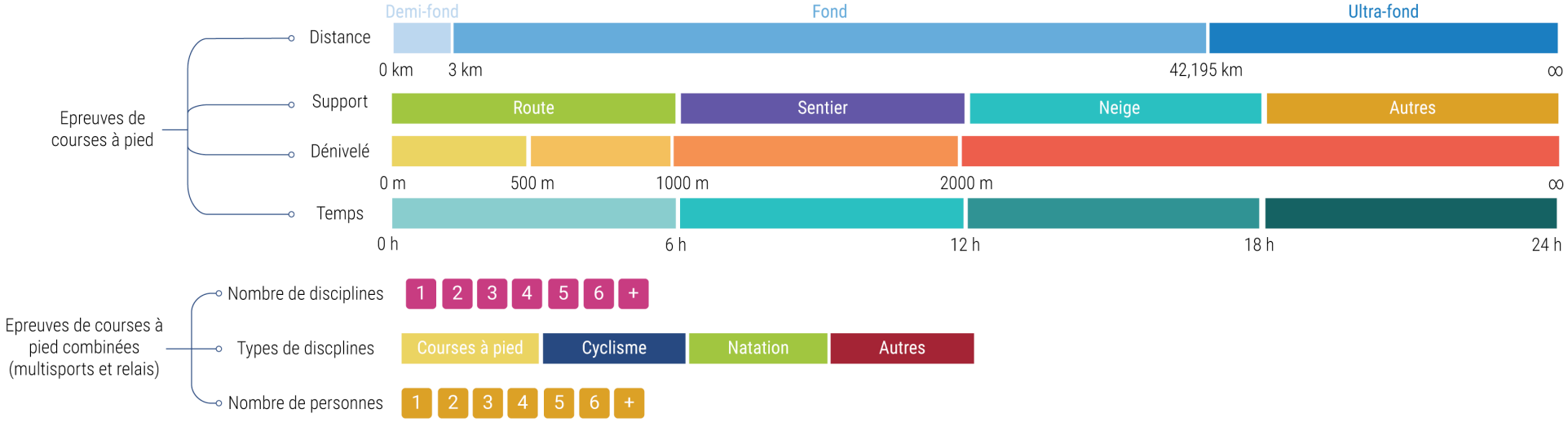
Site-source	Nombre d'attributs	Nombre de courses	Détail des attributs
ACTIVE[5]	23	5725	5K, Ultra, NR, Duathlon, 5 Mile, 1 Mile, Half marathon, 10K, Marathon, Triathlon Sprint, OlympicInternational, 15K, 1K, 8K, Super sprint, Triathlon, 25 Mile, Half century, Racing, Triathlon Half ironman, Ironman, Triathlon OlympicInternational, Triathlon Super sprint, IAAF World Half Marathon Championships
IAAF	32	108	IAAF World Half Marathon Championships, African Championships, Asian Junior Championships, Asian Race Walking Championships, Commonwealth Games, Diamond League Meetings, EAA Outdoor Meetings, IAAF World Relays, European Championships, Games, African Cross Country Championships, IAAF Bronze Label Road Races, IAAF Permit Race Walking Meeting, IAAF Gold Label Road Races, IAAF World Challenge Meetings, IAAF World Championships in Athletics, IAAF World Combined Events Challenge, IAAF World Cup, IAAF World Indoor Championships, IAAF World Race Walking Cup, IAAF World U20 Championships, NACAC Championships, IAAF Silver Label Road Races, Oceania Championships, Oceania Road Championships, Open National Championships, Pan American Championships, South American Championship, South American Road Championships, Youth Olympic Games, IAAF World Cross Country Championships, Oceania Cross Country Championships
Jogging-plus	73	312	8, 11, 20 km, Semi, 3 x 7 km, 10 km, 5, 12, 18 km, 33 km, 10.3 km, 6, 30 km, NR, 11 km, 29 km, 9, 17 km, 35 km, 13 km, 25 km, 50 km, 10.2 km, 5 km, 26 km, 21 km, 10, 16, Marathon, 56 km, 14.7 km, 24 km, 15 km, 7.5 km, 12 km, 42 km, Marathon de Vannes, 23 km, 27 km, 19 km, 7 km, 3 km, 16 km, 9 km, 13, 3, 6 km, Trail urbain 10, 8 km, Trail urbain 5, 31 km, 21, Duo 3.5 km, 45 km, 7, 28 km, 53 km, 10.5 km, 11.6 km, Relais 5+10 km, 32 km, 10.8 km, 12km, 40 km, 14 km, 38 km, 5.1, 13.3 km, 37 km, 22 km, 9.8 km, 10.6 km, 23, Half Marathon
JustRunLah	219	1148	Half Marathon, 10 km, 5 km, 70 m to 3200 m run on grass, 3 km, 6 km & 9 km, 2.5 km, 60 km, 30 km, 23 km (Tough Hilly course), Half (Hilly for 21.1 km), 10.2 km and 3.7 km (Family pair race), 11 km, 22 km, 33 km, 55 km & 80 km, 7 km, 18 km, 10km-60km-10km; 5km-30km-5km, 3 km and 1 km, 20 km, Relay & 64.4 km, 75 km, 43 km, 14 km, 6.5km, 35 laps, 103km, NR, 9 km, 23.5km, 2.4km, 6 km (20-23 Obstacles), 1 km, 2 km, 1.5 km, 35 km, 45 km, 17 km, 5.5 km, Full Marathon, Kids Distance, Sprint Distance, Olympic Distance, 70 km, 21 km, Triathlon, 12 hours, 6 hours, 4 hours, 2.2 km, 1.8 km, 13+ km, 4 km, 15 km, 6 km, 13.5 km, Obstacle Challenge, 25 km and 50 km, 12 hours & 6 hours * Run 2km each loop, 12 km, 1.2 km, 3.5 km, 168 km, 84 km, 50 km, 3 Hour, 6 Hour, 12 Hour, Kids Run, 25 km & 50 km, 48 km and 17 km trail race, Group relay race for 3 hours, Family race for 400 m, and kids run for 1 km, zumba, 12 hours (as many loops as possible), 10 miles, 100 km, 13km, 55 km, Obstacles, Various, 23.54 km on trail course, 25 km, 12.5 km, 13 km, 10 mile, 19 km, 31.6 km, 28 km, 5.6 km, 200km, 71 km, 42 km, 80K Solo, 4-man Relay, 5.7 km, Junior Dash, 23 km, 8 km, 12 hour challenge, 12.9 km, 25.8 km, 64.5 km, 1.2 km Kids Dash, 800 m Kids Dash, 32.1 km trailrace, 53 km, 22.5 km, Relay, 39 km, 5.4 km, 540 m, 36 km, 6.3 km, 27 km, 800 m, 12 km with 4 legs or 8 km with 4 legs as Ekiden relay race and 1 km family race, 32 km, 1.6 km, 25 km on trail, Ironman, 2 km swim, 90 km bike and 21 km run, 84 km scenic

[5] : https://www.active.com/ (consulté le 09/04/2018)

			ride, 75km, 5.25 km, 10.5 km, 11K, 30K, 60K, 9.4 km, 5.1 km, 3 km and 2 km, 8.4 km (10 obstacles), 7 km + 15 Obstacles, 105 km, 50km, 30km, 15km, 444 km, 238 km, Marathon, 31KM, 18KM, 60km, 40km, Ekiden relay race, 150 km, 40 km, 26 km, 64 km, 7km, 150km, 1 lap around the lake for 13.6 km, 70.3 Ironman, 100km, 25K, 50K, 100K, 100K Ekiden, Fun Run, Walk for 5 km and 3 km, 100 miles, 18.45 km, 78 km trail race, 31.7 km, 14.1 km trail races, 2058 steps, 2.5K (Kid Dash), 16.5 km, Relay for 100 km, 1/8 Marathon, 38.55 km, 22.86 km, 10.8 km, 83 km, 3.05 km X 5 = 15.25 km, Run, 19.8 km, Family race, 10.3 km, 0.5 km, 65 km, 8.5 km, 100 km under 13 hours and 50 km under 8 hours, Three-stage Endurance Event (5 km+42.19 5km+8 hours endurance), 1500 m, 44.24 km, 24 hours Relay of over 6 people plus a supporter, 100 mile, 100 km in trail, 65 km individual race, relay race, 80 km, Trail Running, 3 km and kids races, 8 km and 6 km for only Men, 4 km is for only Women, 4.2 km and Ekiden Relay race, 11.5 km fun race, 20 km Relay race, and Road races for 15 km, 45 km Trail race or relay race by 4 people, 20 km and 13 km trail race, 5 hours Endurance Run, 20 km and 7 km, and 2 km for kids, 15 km relay race, 2 km and 300 m for kids, 9.5 km, 16 km, 71.5 km, 10000 m, 5000 m, 400 m, 200 m, 100 m, 5 km
Marathon Ahotu	12	53534	5 km, < 10 km, 10 km, Ultramarathon, 10 km au Semi-Marathon, Semi-Marathon, Course à Etapes, Marathon, Course sans distance, Semi-Marathon au Marathon, Course de X-heures, Course verticale, < 10 km

Source : Projet Running DataLab – Guichet, Plard, 2018

Figure 3. Critères de définition des courses à pied



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