



Table Ronde 10 - les Data Papers

une voie de publication supplémentaire & une source de donnée fiable

Sarah Zahreddine
 Coordinatrice de la gestion des données de recherche
 Équipe de support GDR: rdm-support@ulb.be

<https://portail.ulb.be/fr/recherche/gestion-des-donnees-de-recherche>

27 Mars, 2023



RDA Plenary Cartoons by Auke Herrema.

Définitions

« Articles de données (data paper) : à la différence d'un article scientifique classique qui exploite, analyse et interprète les données scientifiques, un article de données décrit finement un/des jeu(x) de données de façon à en faciliter la compréhension et l'éventuelle réutilisation. »

Glossaire du PNSO2

« Les data papers sont des articles à part entière suivant le même processus éditorial que les articles scientifiques. Ils ont pour but de rendre des jeux de données accessibles, interprétables et réutilisables. »

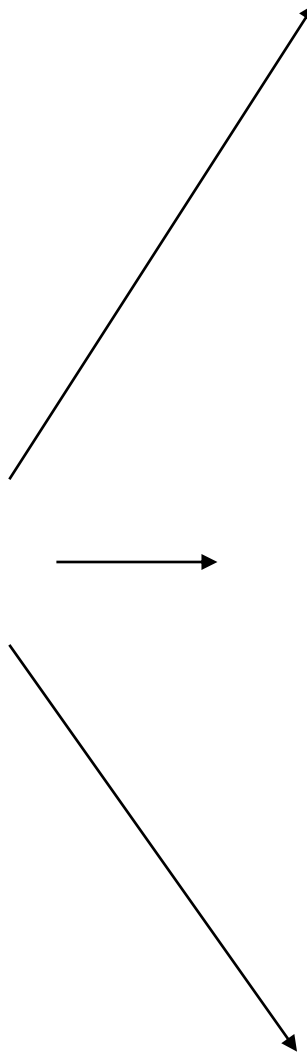
Site de Doranum (Juin 2022)

Data paper, Data Article, Data descriptor, Data Briefs, Resource Announcements, Data Resource Profile

Contenu

- Description des données et contexte
- Description de la méthodologie
 - Protocole, questionnaire, procédure de contrôle de qualité, etc.
- ***Pas de résultats, ni analyses, ni interprétations***
- Montrer le potentiel de réutilisation des données
 - Montrer l'originalité et l'impact sociétale
- Lien vers le jeu de données (préférence dans un dépôt de données)
 - « *As open as possible, as close as necessary* »

Data paper vs



- Section dédiée d'un article scientifique classique (résultats/méthodologie) + référence au jeu de données
- Abstract ou readme d'un dépôt de données
- DMP (Data Management Plan, Plans de gestion de Données)

Objectifs

- Réutilisation des données /reproductibilité des études
 - Documentation complète nécessaire! (ex. Protocoles, questionnaires, readme, métadonnées, etc.)
- Intégrité scientifique : Fiabilité, transparence, qualité des recherche
 - Meilleure garantie contre la fraude scientifique
 - Objectifs de la science ouverte

Exemple

Data in brief 25 (2019) 104344



Contents lists available at ScienceDirect

Data in brief

journal homepage: www.elsevier.com/locate/dib



Data Article

Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico

Fabio Mendoza Palechor*, Alexis de la Hoz Manotas

Universidad de la Costa, CUC, Colombia



ARTICLE INFO

Article history:

Received 3 May 2019
Received in revised form 23 July 2019
Accepted 25 July 2019
Available online 2 August 2019

Keywords:

Obesity
Data mining
Weka
SMOTE

ABSTRACT

This paper presents data for the estimation of obesity levels in individuals from the countries of Mexico, Peru and Colombia, based on their eating habits and physical condition. The data contains 17 attributes and 2111 records, the records are labeled with the class variable NObesity (Obesity Level), that allows classification of the data using the values of Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III. 77% of the data was generated synthetically using the Weka tool and the SMOTE filter, 23% of the data was collected directly from users through a web platform. This data can be used to generate intelligent computational tools to identify the obesity level of an individual and to build recommender systems that monitor obesity levels. For discussion and more information of the dataset creation, please refer

Value of the data

- This data presents information from different locations such as Mexico, Peru and Colombia, can be used to build estimation of the obesity levels based on the nutritional behavior of several regions.
- The data can be used for estimation of the obesity level of individuals using seven categories, allowing a detailed analysis of the affection level of an individual.
- The structure and amount of data can be used for different tasks in data mining such as: classification, prediction, segmentation and association.
- The data can be used to build software tools for estimation of obesity levels.
- The data can validate the impact of several factors that propitiate the apparition of obesity problems.

1. Data

This paper contains data for the estimation of obesity levels in people from the countries of Mexico, Peru and Colombia, with ages between 14 and 61 and diverse eating habits and physical condition as mentioned in [1], data was collected using a web platform with a survey (see Table 1) where anonymous users answered each question, then the information was processed obtaining 17 attributes and 2111 records, after a balancing process described in Figs. 1 and 2. The attributes related with eating habits are: Frequent consumption of high caloric food (FAVC), Frequency of consumption of vegetables (FCVC), Number of main meals (NCP), Consumption of food between meals (CAEC), Consumption of water daily (CH20), and Consumption of alcohol (CALC). The attributes related with the physical condition are: Calories consumption monitoring (SCC), Physical activity frequency (FAF), Time using technology devices (TUE), Transportation used (MTRANS), other variables obtained were: Gender, Age, Height and Weight. Finally, all data was labeled and the class variable NObesity was created with the values of: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III, based on Equation (1) and information from WHO and Mexican Normativity. The data contains numerical data and continuous data, so it can be used for analysis based on algorithms of classification, prediction, segmentation and association. Data is available in CSV format and ARFF format to be used with the Weka tool.

2. Experimental design, materials, and methods

The initial recollection of information was made through a web page using a survey where users had evaluated their eating habits and some aspects that helped to identify their physical condition. The survey was accesible online for 30 days. In Table 1, the questions of the survey are presented.

After all data was collected, then data was preprocessed, so it could be used for different techniques of data mining. The number of records was 485 records, and the data was labeled using equation (1).

Figures/tableaux statistiques descriptives

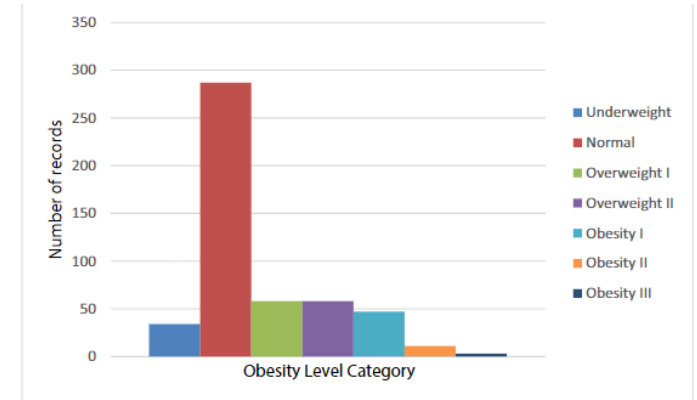


Fig. 1. Unbalanced distribution of data regarding the obesity levels category.

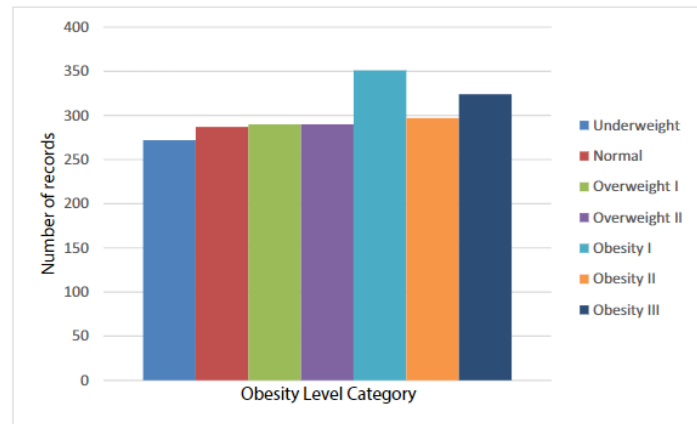


Fig. 2. Balanced Distribution of data regarding the obesity levels category.

Comment trouver un Data Paper?

- Consulter sa communauté scientifique (revues+ littérature citée)
- Scopus/WoS → #Data #Dataset
- Listes disponibles: exemples:
 - <https://doi.org/10.18167/coopist/0057>
 - https://www.forschungsdaten.org/index.php/Data_Journals
 - <https://www.wiki.ed.ac.uk/display/dashare/Sources+of+dataset+peer+review>

Choisir la revue

- Multidisciplinaire vs thématique
- Open Access ou Non
- Rédiger selon le modèle fourni par la revue
- Voir exigence de la revue (comme par ex dépôt de données qui recommandent, modalité d'accès aux données, frais)
- Data journaux (publient uniquement les data papers) vs Revues Classiques (publient articles Classiques + data papers)

Avantages/Bénéfices

- Compte comme un “vrai” papier (peer-review et citable)
- Double visibilité / double avantage / Double chance d’être :
 - visible (indexé dans les bases de données)
 - Trouvé: 2 accès par moteurs de recherche
 - Cité: 2 identifiant numériques, (texte et données)

→ Donc meilleure mise en valeur des données → FAIR DATA

Principes FAIR

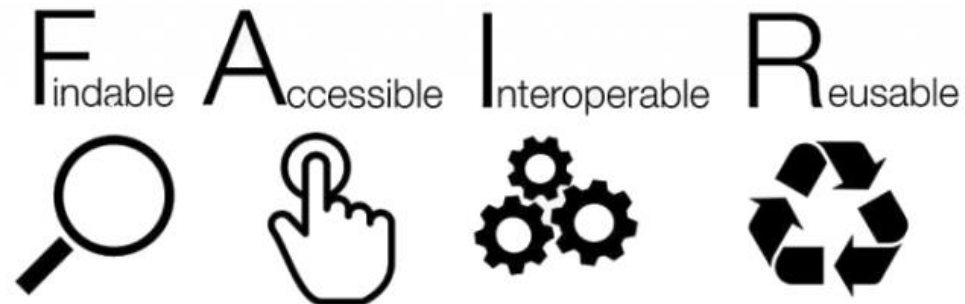
Findable (trouvable) - métadonnées riches, mots clés, identifiant permanent (par ex. DOI)

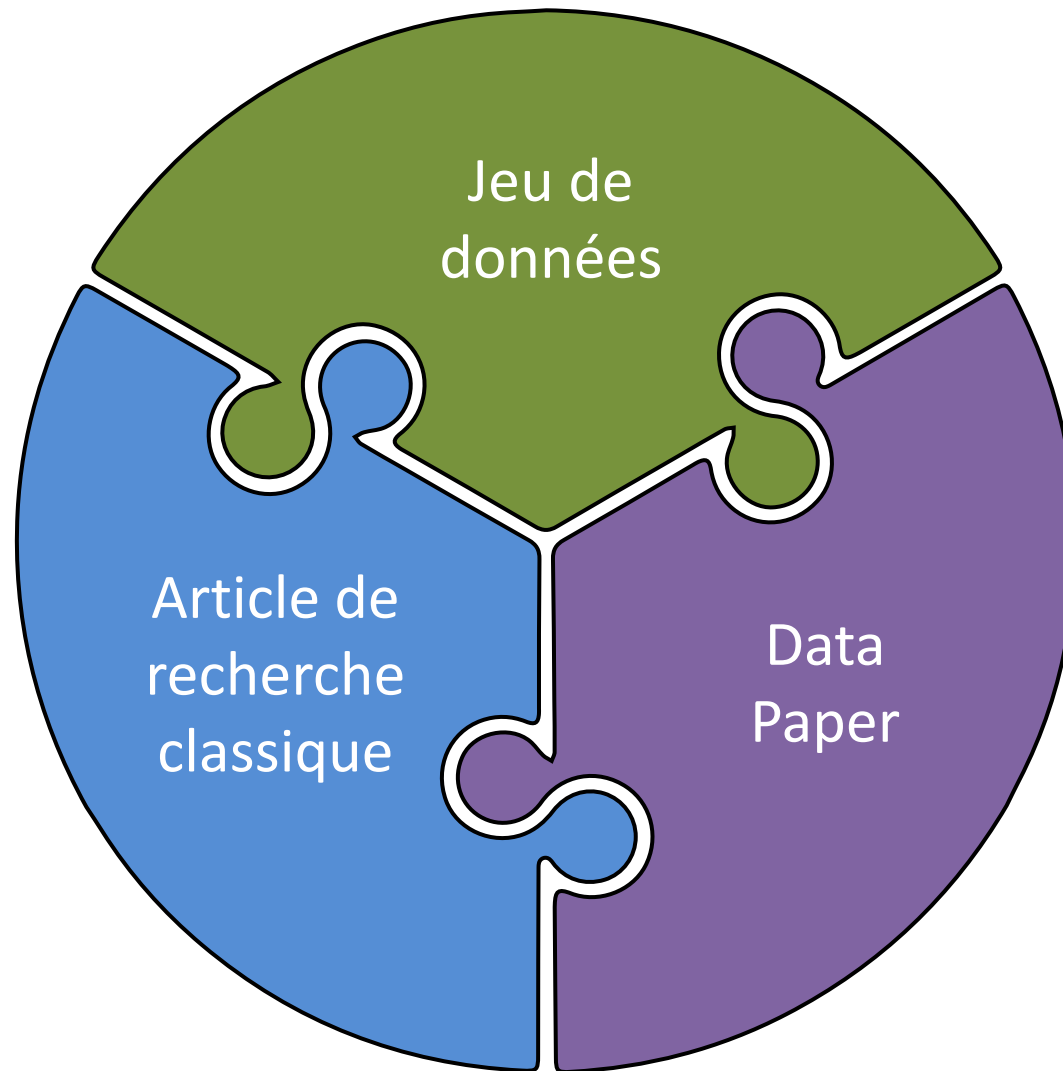
Accessible - dépôt de données fiable, conditions d'accès

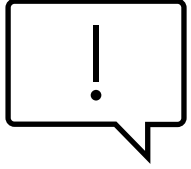
“Accessible/ouvert autant que possible et fermé autant que nécessaire »

Interoperable - appliquer des normes reconnues pour faciliter les échanges et les combinaisons de données

Re-usable (réutilisable) - décrit, documenté, sous licence ([creative commons](https://creativecommons.org/))







Remarques

Avant de divulguer et de partager des données:

- Avec un **potentiel de valorisation et de commercialisation**,
- **Protégées par des droits de propriété intellectuelle** (par exemple, brevets, droits d'auteur, droits de reproduction, etc),
- Qui sont le résultat d'un **projet de recherche en collaboration**,

Le **Technology Transfer Office** de l'ULB (ULB-TTO) doit être consulté (email: ulbtto@ulb.be)

Merci pour votre attention



Questions ?

Sarah Zahreddine

Coordinatrice de la gestion des données de recherche

Équipe de support GDR: rdm-support@ulb.be

<https://portail.ulb.be/fr/recherche/gestion-des-donnees-de-recherche>

27 Mars, 2023