Rédaction, soumission, peer-review : un retour d'expérience multi-perspectives de la publication d'un data paper

GTSO
5 juillet 2022

Frédérique BORDIGNON
École des Ponts & Laboratoire LISIS,
Marne-la-Vallée, France

frederique.bordignon@enpc.fr

Discours & pratiques

3 rôles,
3 points de vue

1. Professionnel de l'IST
   - Documentaliste, data librarian
2. Chercheur
   - Auteur
3. Chercheur
   - Reviewer

Vous avez pensé à publier un data paper ?

Professionnel de l'IST

De quoi parle-t-on ?

"Data papers have been defined as scholarly journal publications whose primary purpose is to describe research data (facts about data). Yet, the literature overview shows that there is a lack of a generally accepted reference definition of data papers."

Schöpfel J. et al. (2019). Data papers as a new form of knowledge organization in the field of research data. *12ème Colloque international d'ISKO-France : Données et mégadonnées ouvertes en SHS : de nouveaux enjeux pour l'état et l'organisation des connaissances ?*, Montpellier, France. ⟨halshs-02284548⟩

"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)

Je préfère faire autrement...

Chercheur

A resituer parmi des alternatives

- Section dédiée d'un article "classique"

  + citation du dataset

A resituer parmi des alternatives

- Section dédiée d'un article "classique"

  + citation du dataset

- Abstract ou readme d'un dépôt de dataset

A resituer parmi des alternatives

Data for: "Evaluation of contaminant retention in the soil of Sustainable Drainage Systems: methodological reflections on the determination of sorption isotherms"

Damien Tedoldi, Kelsey Flanagan, Julien Le Roux, Mohamed Saad, and Marie-Christine Gromaire

LEESU, École des Ponts, UPEC, UPE, Champs-sur-Marne, 6-8 avenue Blaise Pascal, Cité Descartes, 77455 Marne-la-Vallée Cedex 2, France. damien.tedoldi@enpc.fr, kelsey.flanagan@enpc.fr, julien.le-roux@u-pec.fr, mohamed.saad@enpc.fr, marie-christine.gromaire@enpc.fr

The reported data has been acquired within a research project investigating the ability of infiltration-based *Sustainable Drainage Systems* to mitigate pollutant fluxes in stormwater runoff. Sorption experiments have been carried out in batch systems, in order to characterize the sorption behavior of three soils towards copper and zinc, and one soil towards four organic micropollutants: Bisphenol A (BPA), 4-*tert*-Octylphenol (OP), 4-Nonylphenol (4NP), and Nonylphenol Ethoxylate (NP10). The soil samples originate from three source-control infiltration facilities; their main characteristics are summarized in Table 1.

- *Soil 1* is a sandy loam from a roadside swale;
- *Soil 2* is a silt loam from a roadside filter strip;
- *Soil 3* is a sandy clay loam from an infiltration basin.

Table 1. Main physical and physicochemical properties of the characterized soils.

	Soil 1	*Soil 2*	*Soil 3*
Texture (USDA classification)	Sandy loam	Silt loam	Sandy clay loam
Sand [%]	72	10	47
Silt [%]	18	67	25
Clay [%]	10	23	28
Bulk density [kg.dm^{-3}]	1.5 ± 0.1	1.2 ± 0.1	1.1 ± 0.1
Saturated water content [cm^3.cm^{-3}]	0.36 ± 0.07	0.44 ± 0.04	0.47 ± 0.05
Organic carbon (f_{OC}) [g.kg^{-1}]	8.7	17.8	41.6
Volatile matter [g.kg^{-1}]	30	51	85
CEC [cmol$_+$.kg^{-1}]	5.1	11.8	13.6
pH$_{water}$	8.7	8.7	8.2
Carbonates [g.kg^{-1}]	240	29	32

All experiments were carried out in non-competitive systems (*i.e.*, individual contaminants in each batch system). The "reference conditions" corresponded to an electrolyte solution which mimicked the composition of runoff water (*Evian* and ultrapure water, volumetric ratio of 1:10). The sorption behavior of metals onto soils 1 and 3 was also investigated in different conditions:

- 1.0 g.L^{-1} of sodium chloride, to investigate the effects of deicing salt in runoff water;
- 10 mg.L^{-1} of humic acids, to represent the natural generation of dissolved organic matter in the soil solution.

The final state of each batch system provided one point of the sorption isotherm, *i.e.* one pair (C_{eq}, S_{eq}) where C_{eq} and S_{eq} are respectively the dissolved concentration [mg.L^{-1}] and solid content [mg.kg^{-1}] at equilibrium. The latter was calculated as $S_{eq} = S_i + \Delta S$, where S_i is the initial content of the soil and ΔS is the sorbed amount of contaminant [mg.kg^{-1}]:

$$\Delta S = \frac{(C_i - C_{eq})V}{M}$$

where C_i is the initial concentration in the batch system [mg.L^{-1}], V is the volume of solution [L], and M is the mass of suspended soil [kg]. The methodology used to select the experimental conditions (C_i and M/V ratio) is thoroughly described in a research article[1].

Data description

The appended files present the experimental points of the sorption isotherms for each characterized soil.

- *Soil 1*: metals in the reference conditions (Metals_ref), metals with sodium chloride (Metals_NaCl), metals with humic acids (Metals_HA), organic micropollutants;
- *Soil 2*: metals in the reference conditions;
- *Soil 3*: metals in the reference conditions, metals with sodium chloride, metals with humic acids.

The concentrations in the dissolved phase are referred to as "X_diss", where X is the studied species (Cu, Zn, *etc.*), and are expressed in mg.L^{-1}. BPA stands for Bisphenol A, OP stands for 4-*tert*-Octylphenol, 4NP stands for 4-Nonylphenol, and NP10 stands for Nonylphenol Ethoxylate. The contents in the solid phase are referred to as "X_sorb" and are expressed in mg.kg^{-1}.

In case the tests were carried out in duplicates, the columns "X_diss" and "X_sorb" correspond to the mean values of the experiments; two additional columns, referred to as "sdX_diss" and "sdX_sorb", provide the standard deviations (also expressed in mg.L^{-1} and mg.kg^{-1}, respectively).

[1] Damien Tedoldi, Kelsey Flanagan, Julien Le Roux, Ghassan Chebbo, Philippe Branchu, Mohamed Saad, and Marie-Christine Gromaire. 2019. Evaluation of contaminant retention in the soil of Sustainable Drainage Systems: methodological reflections on the determination of sorption isotherms. *Blue-Green Systems* 1(1). doi: 10.2166/bgs.2019.196

f soil contamination in

enance operations"

nas, M.-C. Gromaire

investigating soil contamination in
(i) the spatial distribution of three
llutants by the US-EPA) within each
nter-site differences in contamination

ance), called *Seine-et-Marne*, which
contains around 4,500 kilometers of
selected across all the territory's area
lings, open countryside, forest), and
ed by (i) a number from 1 to 40, and
here "RD" stands for "departmental
x was completed by the letters "A",
and by two road factors which were
traffic and the percentage of semi-

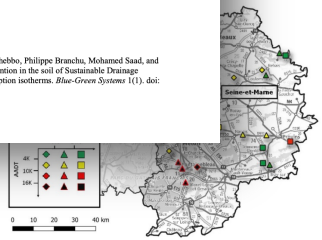


Fig. 1 – Location of the study sites in the Paris region. The colors and shapes indicate respectively the classes of traffic and the percentage of trucks on the adjacent road.

At each site, soil samples were collected at four to five distances from the road, depending on the shoulder width (0, 0.3, 0.7, 1.2, and 1.8-3 m whenever possible). For each distance, a composite sample was formed from 6 evenly spaced points along a transect parallel to the road. The targeted depth corresponded to the upper 5 centimeters, however in some cases the presence of the road sub-base only allowed the collection of a surface sample (0-2 cm). Due to cost constraints, only two samples from each site could be analyzed for PAHs: it was chosen to retain (i) a composite sample formed by merging subsamples collected at 0 and 0.3 m, and (ii) a sample collected on the opposite side from the road.

The soil samples intended for **metal analysis** were pre-treated according to the standard ISO 11464[1]. Analysis was then achieved *via* X-ray fluorescence spectrometry (*Thermo Scientific*, portable Niton™ analyzer XL3t in a standard soil mode, 60 s per beam). 5 to 6 homogenized subsamples were analyzed independently, then each sample was assigned its average concentrations of copper, lead and zinc, which are reported in the present dataset.

The samples intended for **PAH analysis** were sent to an analytical laboratory. The analyzed compounds were the 16 PAHs classified as priority pollutants by the US-EPA[2]. After fine grinding of the samples, PAHs were extracted with a mixture of acetone and dichloromethane, followed by GC/MS analysis. The complete method is referenced in the European standard EN 15527[3].

Data description

The table **Site characteristics** presents, for each site:

- the site index from 1 to 40 (column **Site**),
- the site index in the French nomenclature (column **RD**),
- the GPS coordinates of the sites (columns **X-coord** and **Y-coord**),
- the annual average daily traffic, expressed in vehicles per day (column **AADT**),
- the truck fraction, expressed in % (column **Trucks**).

The tables **Metals** and **PAHs** present the concentrations analyzed in each collected sample. These two tables have a similar structure, with:

- the site indexes (columns **Site** and **RD**),
- the sampling distance, expressed in meter from the road (column **Distance**),
- the soil concentrations of the analyzed contaminants, expressed in milligram per kilogram of dry matter:
  - **Metals**: copper (column **Cu**), lead (column **Pb**), zinc (column **Zn**);
  - **PAHs**: full name of the 16 analyzed compounds[2,4].

[1] ISO 11464 (2006) Soil quality – Pretreatment of samples for physico-chemical analysis. International Organization for Standardization, Geneva.

[2] Naphthalene, acenaphthylene, acenaphthene, fluorene, phenanthrene, anthracene, fluoranthene, pyrene, benzo(a)anthracene, chrysene, benzo(b)fluoranthene, benzo(k)fluoranthene, benzo(a)pyrene, dibenzo(ah)anthracene, indeno(123,cd)pyrene, and benzo(ghi)perylene.

[3] EN 15527 (2008) Characterization of waste – Determination of polycyclic aromatic hydrocarbons (PAH) in waste using gas chromatography mass spectrometry (GC/MS). European Committee for Standardization, Brussels.

[4] Results are not available for site #4 (RD 51) due to a problem during sample transport.

Abstracts/readme de dépôts sur Mendeley Data

A resituer parmi des alternatives

- Section dédiée d'un article "classique"

  + citation du dataset

- Abstract ou readme d'un dépôt de dataset

- Article exécutable

Jonathan Zurbach, Univ Avignon

Journal of Digital History - https://journalofdigitalhistory.org/en

Observable - https://observablehq.com/

A resituer parmi des alternatives

- Section dédiée d'un article "classique"

  + citation du dataset

- Abstract ou readme d'un dépôt de dataset

- Article exécutable

Jonathan Zurbach, Univ Avignon
Journal of Digital History - https://journalofdigitalhistory.org/en
Observable - https://observablehq.com/

- DMP

Publier un data paper, mais où ?

Chercheur

Où sont vraiment publiés les data papers ?

Cherchons les…

- Scopus : 10 000+ data papers sur 80+ millions de résultats

TITLE-ABS-KEY(data OR dataset OR database OR "the" OR "is" OR "are" OR "of" OR "and" OR "have" OR "has")
AND (LIMIT-TO (DOCTYPE,"dp"))

- WoS : 12 000+ data papers sur 65+ millions de résultats

(TS=((data OR dataset OR database OR "the" OR "is" OR "are" OR "of" OR "and" OR "have" OR "has")))
AND (DT==("DATA PAPER"))

- Dimensions : ?
- The Lens : ?
- Summon : ?
- Isidore : ?
- HAL : ?

Où sont vraiment publiés les data papers ?

Cherchons les bien…

- Scopus : 117 000+ résultats

(TITLE ((dataset OR "data of" OR "data on" OR "data from" OR "survey data" OR "supplementary data" OR "data about")) AND NOT TITLE ("based on" OR "using"))

- WoS : 104 000+ résultats

TI=((dataset OR "data of" OR "data on" OR "data from" OR "survey data" OR "supplementary data" OR "data about")) NOT TI=(("based on" OR "using"))

- Dimensions : 165 000+ résultats

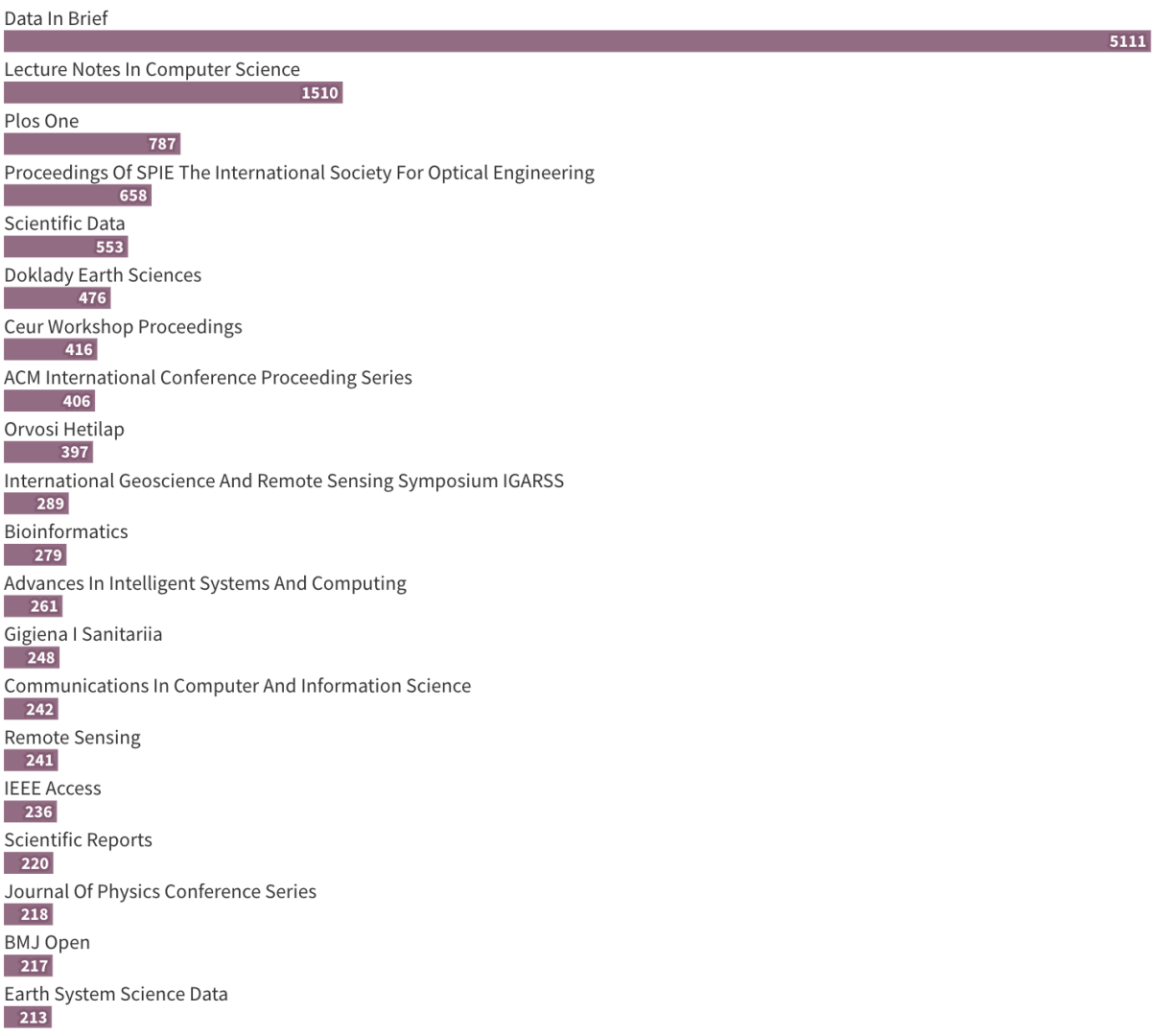
title:((dataset OR "data of" OR "data on" OR "data from" OR "survey data" OR "supplementary data" OR "data about") NOT ("based on" OR "using"))

- The Lens : 1 800 000+ résultats
- Summon : 8 400 000+ résultats
- Isidore : 13 résultats
- HAL : 3 500+ résultats

Nombre de
pseudo data papers
par revue

Top 20

(pseudo requête)



Un data paper,

c'est
ça compte comme
} un vrai papier

Professionnel de l'IST & Chercheur

Peer-reviewed

Citable

Ambiguïté du type

ELSEVIER

Data in Brief

Volume 34, February 2021, 106731

Data Article

Dataset of search queries to map scientific publications to the UN sustainable development goals

Frederique Bordignon # ✉ ⊕

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.dib.2021.106731

Data in Brief...

Chercheur

L'expérience *Data in Brief*

- Pre-review


Title: it must include the word 'data' or 'dataset'.


Données d'enquête -> uploader le questionnaire à part

In the 'Ethics' section we require a statement to confirm that informed consent of all participants has been obtained. Please indicate if this is the case for your study in the Ethics section.


Données scrapées


URL secrète de Mendeley Data -> inutile

- Review

- Pre-rev

It mus

Donné

In the 'Ethics' sec

Please indicate if

Donné

URL s

- Review

Specifications Table

Subject	Library and Information Sciences
Specific subject area	Information retrieval, bibliometrics
Type of data	Search strings in txt files Tabular data in a csv file
How data were acquired	Search strings: manually created Survey data: collected with LimeSurvey
Data format	Raw, txt and csv files
Parameters for data collection	In order to mitigate the polysemy of terms as much as possible, we identified and targeted the most relevant subject areas for each SDG. In addition, we also used a text-mining tool (CorTexT) to identify as many relevant phrases as possible. The results were submitted (through LimeSurvey) to researchers for their feedback.
Description of data collection	These data were manually written in a text editor.
Data source location	Ecole des Ponts Marne-la-Vallée France
Data accessibility	Repository name: Mendeley Data Data identification number: http://dx.doi.org/10.17632/xrx7ddbbb4.1 Direct URL to data: http://dx.doi.org/10.17632/xrx7ddbbb4.1

s has been obtained.

L'expérience *Data in Brief*

- Pre-review


It must include the word 'data' or 'dataset'.


Données d'enquête -> uploader le questionnaire à part -> exercice

In the 'Ethics' section we require a statement to confirm that informed consent of all participants has been obtained. Please indicate if this is the case for your study in the Ethics section.


Données scrapées


URL secrète de Mendeley Data

- Review -> Specifications table
- Publication -> Proofs & coquilles, lien fort avec le dépôt
- HAL...

Evaluer un data paper, ou évaluer un dataset ?

Reviewer

Evaluer un data paper, évaluer les données…

Reviewer Recommendation Term:	Major Revision	
Manuscript Rating Question(s):	**Scale**	**Rating**
*1) Are these data original and produced by the authors? [1=Yes, 2=No, 3=N/A]	[1-3]	2
*2) Are these data secondary (e.g. censuses, government databases, organizational records)? [1=Yes, 2=No, 3=N/A] If YES, please answer 2a, 2b & 2c; if NO go to 3	[1-3]	1
2a) Secondary Data Only: were these data collected using variables that make the study unique?	[1-3]	N/A
2b) Secondary Data Only: is this collection of secondary data of value to the research community?	[1-3]	1
2c) Secondary Data Only: do the authors provide the protocol for collecting/creating these data?	[1-3]	1
*3) Have the authors used a questionnaire or survey? [1=Yes, 2=No, 3=N/A] If YES, please answer 3a, 3b, 3c; if NO go to 4	[1-3]	3
3a) Questionnaire/survey only: is the questionnaire/survey direct, unambiguous, unbiased?	[1-3]	N/A
3b) Questionnaire/survey only: is the sampling representative of the population and rigorously following a scientific method?	[1-3]	N/A
3c) Questionnaire/survey only: have the authors provided the full questionnaire/survey used?	[1-3]	N/A
*4) Do the authors adequately explain to the research community the utility of these data in the "Value of data" section? [1=Yes, 2=No, 3=N/A]	[1-3]	1
*5) Are these data described clearly in the "Data" section? [1=Yes, 2=No, 3=N/A]	[1-3]	2
*6) Is the protocol/method for generating these data adequately described in "Experimental design, materials, and methods" section? [1=Yes, 2=No, 3=N/A]	[1-3]	2
*7) Have the authors provided the raw data related to any tables, graphs, images and charts, etc.? [1=Yes, 2=No, 3=N/A]	[1-3]	3
*8) Are these data publicly available (they may be hosted with the article or deposited externally in a repository)? [1=Yes, 2=No, 3=N/A]	[1-3]	1
*9) If these data are deposited in a repository, is there a link to the data and does it work? [1=Yes, 2=No, 3=N/A]	[1-3]	1
*10) If this data article is related to an existing primary research article is there any duplication? If yes, please comment on this. [1=Yes, 2=No, 3=N/A]	[1-3]	N/A

- Adéquation entre la description des données et les données associées

  Contenu, version

- Questions juridiques : sur les données ? Ou uniquement le papier ?

Merci

Questions ?

Frédérique BORDIGNON
École des Ponts & Laboratoire LISIS

frederique.bordignon@enpc.fr

slidesgo