Data publishing

Joachim Schöpfel

Data publishing

Joachim Schöpfel

Maître de conférences, GERiiCO, Université de Lille 3

joachim.schopfel@univ-lille3.fr

https://fr.linkedin.com/in/schopfel

Vous publiez ? Oubliez les articles. Dépassés. Trop longs. Trop lents. Trop consommateur de temps. En plus, la moitié des manuscrits sont rejetés. Et même si votre article est accepté – sérieusement, qui aura le temps de le lire ? Tout le monde vous le dira : il y a trop de revues, il y a trop d'articles, tout cela coûte cher, la qualité n'est pas toujours au rendez-vous ; et ne parlons pas des failles du facteur d'impact.

Oubliez donc les articles. Ce qui compte aujourd'hui, ce sont les données. Vous avez entendu parler du *Open Data*, de l'ouverture des données publiques. Vous avez du entendre promouvoir le *Big Data* comme nouveau challenge et eldorado de l'économie numérique. Et à coup sûr, quelqu'un – votre VP recherche, votre directeur IST, votre Ministère, la Commission Européenne – vous a fait remarquer récemment qu'il faut mettre les résultats bruts à la disposition de tout le monde dans la mesure où votre recherche est payée par les contribuables. Et ceci dans un régime aussi libéral que possible, pour stimuler la recherche et l'économie. *Open science* oblige.

Les précurseurs de l'eScience l'avaient annoncé : la science de l'avenir est faite de données, pas de publications (1). Les uns, plus sceptiques, affirmaient que les données allaient juste compléter le rôle traditionnel des publications (2). Les autres, visionnaires, pressentirent qu'elles se substitueraient aux revues et aux livres : "In the age of genomic-sized datasets, the biomedical literature is increasingly archaic as a form of transmission of scientific knowledge for computers" (3). Certes, le 4e paradigme de la découverte scientifique (4) ne transformera pas toutes les disciplines en même temps et au même rythme, et certains domaines sont plus exposés que d'autres. Mais même les sciences humaines et les arts s'y mettent aujourd'hui.

Archaïque, la littérature scientifique – le ton est donné. Qui a envie d'être archaïque ? Oubliez donc les articles et publiez vos données. Mais comment ? Publier ses données, cela ne s'invente pas. Il faut y penser dès le début – l'organisation des données, leur description, leur mise en forme et présentation, leur lisibilité, leur format et les éventuels problèmes de droit (5). D'autres questions se poseront par la suite : où faut-il publier les données ? Vous pouvez déposer vos données dans une archive de données. Il y en a dans toutes les disciplines. Le répertoire re3data.org recense plus de 1300 dont 64 en France ; HAL à Lyon se positionne comme future archive de données nationale. Mais vous pouvez aussi préparer un petit *data paper* pour l'une des nouvelles « revues de données » dont le nombre ne cesse de croître, comme par exemple *Data in Brief*, *Genomics Data*, *Scientific Data*, *GigaScience* ou *Research Data Journal for the Humanities and Social Sciences*.

En fait, ces *data papers* sont un compromis avec les articles traditionnels car ils contiennent, outre une description précise des données, une analyse avec discussion des résultats. Mais ils sont bien plus courts, souvent seulement deux à trois pages. Et très bien structurés, c'est-à-dire « machine-

readable » et exploitables par des outils du *text and data mining* (TDM). Publier la science, oui, mais sous forme de données. C'est l'avenir de la science, et cet avenir commence sous nos yeux.

PS. Vous pensez que j'exagère ? Vous avez raison. Quoique…

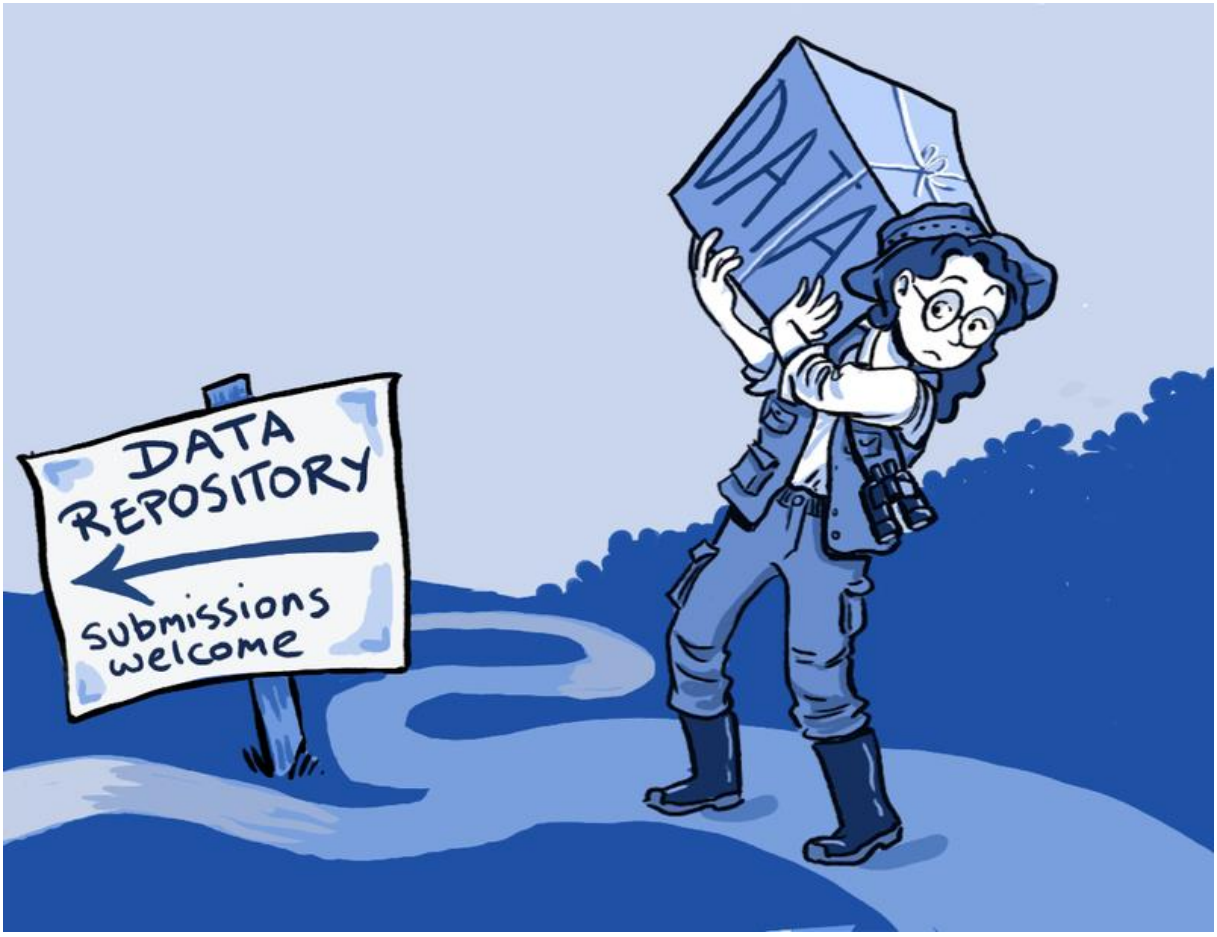
(1) Hey, T., Trefethen, A. E., 2005. Cyberinfrastructure for e-Science. *Science* 308 (5723), 817-821. URL http://dx.doi.org/10.1126/science.1110410

(2) Borgman, C. L., Wallis, J. C., Enyedy, N., 2007. Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries* 7 (1-2), 17-30. http://escholarship.org/uc/item/6fs4559s

(3) Blake, J. A., Bult, C. J., 2006. Beyond the data deluge: Data integration and bio-ontologies. *Journal of Biomedical Informatics* 39, 314-320. http://dx.doi.org/10.1016/j.jbi.2006.01.003

(4) Hey, T., Tansley, S., Tolle, K. (dir.), 2009. The fourth paradigm. Data-intensive scientific discovery. Microsoft Corporation, Redmond, WA. http://research.microsoft.com/en-us/collaboration/fourthparadigm/

(5) Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., Dorsett, K., Aug. 2015. Changes in data sharing and data reuse practices and perceptions among scientists worldwide. PLoS ONE 10 (8), e0134826+. http://dx.doi.org/10.1371/journal.pone.0134826



"To deposit or not to deposit, that is the question - journal.pbio.1001779.g001" by Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, et al. (2014) - Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, et al. (2014) Troubleshooting Public Data Archiving: Suggestions to Increase Participation. PLoS Biol 12(1): e1001779. doi:10.1371/journal.pbio.1001779. Licensed under CC BY 4.0 via Commons - https://commons.wikimedia.org/wiki/File:To_deposit_or_not_to_deposit,_that_is_the_question_-_journal.pbio.1001779.g001.png#/media/File:To_deposit_or_not_to_deposit,_that_is_the_question_-_journal.pbio.1001779.g001.png