BLACK FAIRDAY

Michael Nauge

▶ **To cite this version:**

Michael Nauge. BLACK FAIRDAY. Humanistica 2020, May 2020, Bordeaux, France. hal-02561825

HAL Id: hal-02561825
https://hal.science/hal-02561825

Submitted on 4 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

BLACK FAIRDAY

Michael Nauge[1,2]

Laboratoire FoReLLIS[1], Laboratoire MIMMOC[2]

Université de Poitiers

Le Black Friday c'est terminé, faites place au Black Fairday et profitez de promotions exceptionnelles sur les documents computationnels. Après la ruée vers l'OR, il faut maintenant se dépêcher de se ruer vers le FAIR. Nouvelle valeur refuge grâce à son fort pouvoir de réussite aux appels à projets[1], il vous garantit des sources de financement pour vos projets de recherche. N'hésitez plus, investissez !

« Oui j'en veux, mais comment faire FAIR avec mon corpus spécifique et complexe ? » me direz-vous. Faites au mieux, il n'y a pas réellement une norme mais plutôt une « cascade de standards »[2] et de recommandations. Tout est encore en construction[3] et se coordonne avec entre autres, l'implication de la TGIR Huma-Num par son pilotage de CO-OPERAS IN[4] qui « a pour objectif d'organiser et de superviser l'implémentation des données de la recherche en sciences humaines et sociales selon les principes de l'initiative GO FAIR (Findable, Accessible, Interoperable and Reusable) ».

En considérant que la TGIR Huma-Num est un acteur incontournable pour le FAIR, nous pouvons nous appuyer sur les services qu'ils proposent. Un moissonnage par le service Isidore peut jouer un rôle important pour le F (Facile à trouver), tandis que Nakala est un bon candidat pour le A (Accessibilité) avec son service de stockage, de description avec des métadonnées standards et l'attribution de Handle/DOI pour chaque ressource. Pour le I (Interopérabilité), elle peut être technique avec les multiples API (Access Point Interface) proposées par Isidores et Nakala. L'interopérabilité peut être aussi sémantique avec l'utilisation d'OpenTheso. C'est pour le R (Réutilisable) qu'il nous semble devoir se questionner réellement.

La réutilisation de corpus implique évidemment le respect des paradigmes F, A et I mais elle requière avant tout : la confiance. Cette dernière est la plus délicate à obtenir en période de crise de la reproductibilité[5]. Il est indispensable de connaître la genèse du corpus, la manière dont les données ont été collectées, décrites, nettoyées, filtrées, transformées, transcodées, enrichies, biaisées et interrogées. Dans cet objectif, la rédaction et le partage d'un plan de gestion de données est une bonne initiative. En complément, un outil très utilisé en science expérimentale est le cahier de laboratoire[6] qui permet une « transmission de savoir » et une grande traçabilité. Cependant ces deux solutions

---

[1] Féret, Romain, Bracco, Laetitia, Cheviron, Stéphanie, Lehoux, Elise, Arènes, Cécile, & Li, Ling. (2020). Améliorer les chances de succès de son projet ANR grâce à la Science Ouverte (Version 1). *Zenodo*. http://doi.org/10.5281/zenodo.3741666 [en ligne]

[2] Schöpfel, J. (2018). Hors norme ? Une approche normative des données de la recherche. *Revue COSSI*, n°5 [en ligne]

[3] Fair data alliance (2020), FAIR Data Maturity Model: specification and guidelines [en ligne]

[4] Busonera, P. (2019). L'ANR finance le réseau CO-OPERAS IN dans le cadre de son appel à projet Science Ouverte sur les données de la recherche [en ligne]

[5] Pierre Carl Langlais et EPRIST (2020), La recherche en crise de reproductibilité ?, *EPRIST Analyse I/IST n°30* [en ligne]

[6] CNRS. Présentation du cahier de laboratoire [en ligne]

semblent faibles pour suivre les mutations constantes des données numériques soumises à un écosystème hostile de logicielles et de scripts en tous genres. Un début de solution apparaît avec la maturité des documents computationnels tels que les Jupyter notebook, R notebook, Org-mode, Lodide qu'il convient d'accompagner d'un système de gestion de version comme GIT[7].

Les documents computationnels (ou carnets de code) permettent d'entrelacer habillement l'édition savante traditionnelle (en langage naturelle et mathématique) et des scripts (en langage de programmation) afin d'expliciter les manipulations opérées sur les données, le tout accompagné de représentations graphiques interactives. Les cas d'usages sont variés, dont le cahier de laboratoire interactif pour le suivi d'un projet au quotidien[8]. Ces documents sont aussi très pratiques en support de cours[9] pour la formation des étudiants aux parcours colorés « SHS data scientist » en facilitant l'initiation à la formalisation de questionnements scientifiques, d'écriture d'algorithme et de code. Enfin le cas d'usage où le document computationnel semble vraiment faire la différence en contexte de crise de reproductibilité, est lorsque l'on publie un article scientifique et qu'en parallèle nous disposons d'un dépôt de données accessible contenant le corpus exploité par la publication. Ce type de document vient combler l'espace existant entre les données et leur interprétation. Il peut rendre transparent la manière dont le corpus a été produit et la manière dont il est interrogé pour générer des vues interprétables pour l'analyse[10]. En rendant disponible ce type de document, l'appropriation et réutilisation d'un corpus devient enfin atteignable en réduisant considérablement le temps nécessaire à l'interrogation dynamique du corpus accessible.

Derrière ses allures sexy pour « techno-addict », le document computationnel a le potentiel pour bouleverser réellement la publication scientifique et la réutilisation de corpus. Plus qu'un outil technique, il semble pouvoir transpercer toutes les disciplines par sa souplesse de manipulation de données de toutes natures. Le frein actuel est lié à une montée en compétence pour l'assemblage de fragments de code d'exploration de données et l'appropriation d'un gestionnaire de version pour que les producteurs de corpus puissent enfin faire vivre leurs données sans peur de perte de traçabilité.

---

[7] Champin, P. H. (2013). Introduction à GIT. [en ligne]

[8] Hanote, S. et al. (2018). DicoDiachro project [notebook en ligne]

[9] Nauge, M. (2019). Marmython : Une recette facile pour cuisiner un corpus au Python. *Exploiter les corpus d'auteur – Atelier de formation annuel du consortium Cahier* [notebook en ligne]

[10] Duchet, J. et al. (2019) Syneresis studies on Buchanan and Walker dictionaries. Complément de la communication : « Les terminaisons -ic et -ical en anglais : essai de comparaison métalexicographique entre le dictionnaire de Buchanan (1766) et de Walker (1791) ». *10èmes Journées Internationales de la Linguistique de corpus.* [notebook en ligne]