Un data paper en SHS
Pourquoi, pour qui, comment ?

Victor Gay

TSE, IAST, UT1

#dhnord2021

Un retour d'expérience de production de data paper

- Une base de données : le TRF-GIS.
- Un SIG pour la France de la Troisième République.

Un retour d'expérience de production de data paper

- Une base de données : le TRF-GIS.
- Un SIG pour la France de la Troisième République.
- Nomenclatures et shapefiles annuels (1870-1940) :
  - Circonscriptions générales : départements, arrondissements, cantons, communes.
  - Circonscriptions spéciales : militaires, judiciaires, pénitentiaires, électorales, académiques, inspections du travail, ecclésiastiques.

Un retour d'expérience de production de data paper

- Une base de données : le TRF-GIS.
- Un SIG pour la France de la Troisième République.
- Nomenclatures et shapefiles annuels (1870-1940) :
  - Circonscriptions générales : départements, arrondissements, cantons, communes.
  - Circonscriptions spéciales : militaires, judiciaires, pénitentiaires, électorales, académiques, inspections du travail, ecclésiastiques.
- 16 bases de données :
  - 901 nomenclatures, 830 shapefiles.
  - Matériel de reproduction : code source, données source, archives.

Un retour d'expérience de production de data paper

- Une base de données : le TRF-GIS.
- Un SIG pour la France de la Troisième République.
- Nomenclatures et shapefiles annuels (1870-1940) :
  - Circonscriptions générales : départements, arrondissements, cantons, communes.
  - Circonscriptions spéciales : militaires, judiciaires, pénitentiaires, électorales, académiques, inspections du travail, ecclésiastiques.
- 16 bases de données :
  - 901 nomenclatures, 830 shapefiles.
  - Matériel de reproduction : code source, données source, archives.
- Entrepôt : https://dataverse.harvard.edu/dataverse/TRF-GIS.

Un retour d'expérience

- Un *data paper* :
  - Gay, Victor. 2021. « Mapping the Third Republic. A Geographic Information System of France (1870-1940) ». *Historical Methods : A Journal of Quantitative and Interdisciplinary History*. À paraître.
  - DOI : 10.1080/01615440.2021.1937421
  - HAL : https://hal.archives-ouvertes.fr/hal-02951461.

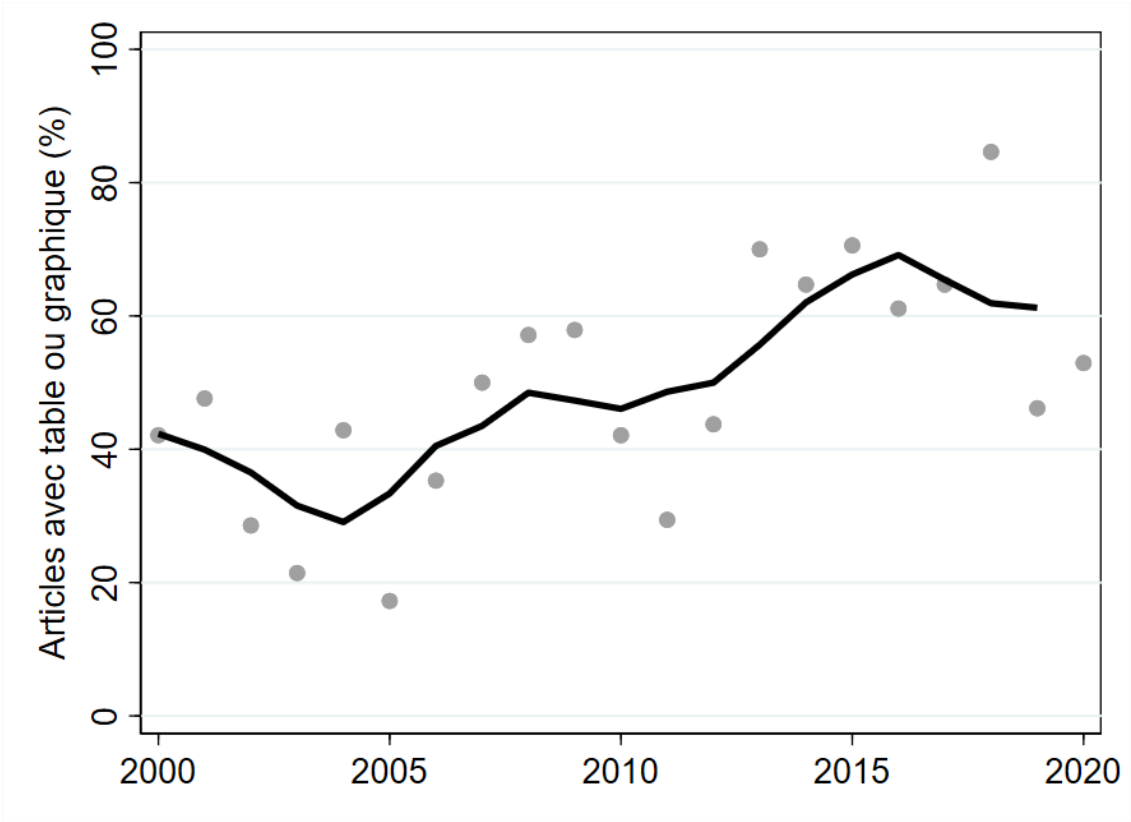
Un *data paper* en SHS : pourquoi ?

Le tournant de la quantification

- Tournant quantitatif en SHS depuis 10 ans.

Le tournant de la quantification

- Tournant quantitatif en SHS depuis 10 ans.
- En sociologie.



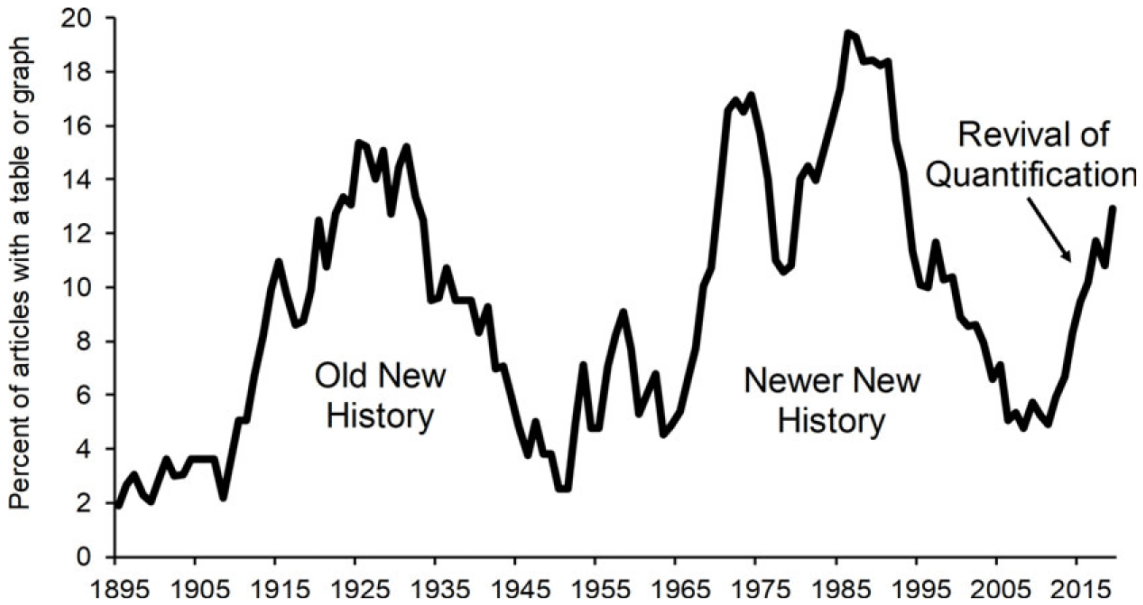
Pourcentage d'articles de la *Revue française de sociologie* avec au moins une table ou graphique de statistiques. Moyenne mobile sur 5 ans. Calculs de l'auteur.

Le tournant de la quantification

- Tournant quantitatif en SHS depuis 10 ans.
- En histoire (française).
  - Numéro « Histoire quantitative » des *Annales. Histoire, sciences sociales*. 2018.
  - Zalc, Claire et Claire Lemercier. 2008. *Méthodes quantitatives pour l'historien*. Paris : La Découverte.
  - Zalc, Claire et Claire Lemercier. 2020. *Quantitative Methods in the Humanities. An Introduction*. Charlottesville (Virg.) : University of Virginia Press.

Le tournant de la quantification

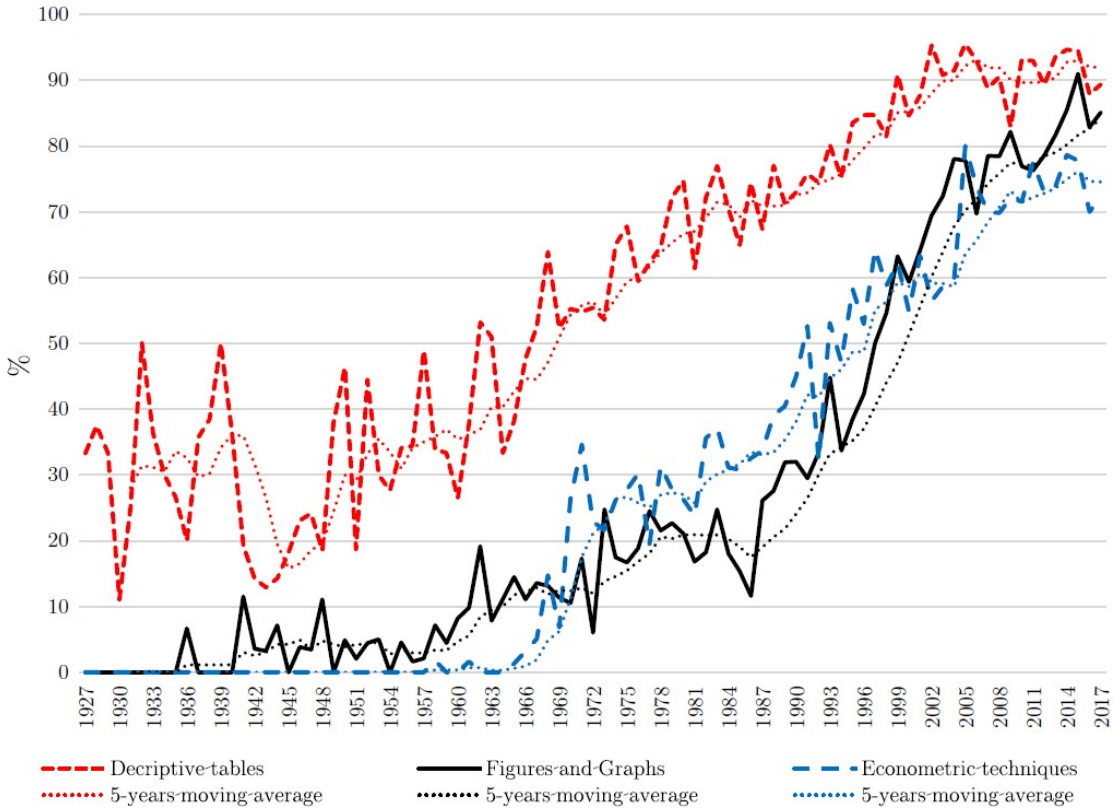
- Tournant quantitatif en SHS depuis 10 ans.
- En histoire (anglo-saxonne).



Pourcentage d'articles de l'*American History Review* avec au moins une table ou graphique de statistiques. Moyenne mobile sur 7 ans. Ruggles (2021, 14)

Le tournant de la quantification

- Tournant quantitatif en SHS depuis 10 ans.
- En histoire économique (anglo-saxonne).



Pourcentage d'articles avec au moins une table ou graphique de statistiques dans le l'*Economic History Review*, le *Journal of Economic History*, *Explorations in Economic History*, la *European Review of Economic History*, et *Cliometrica*. Moyenne mobile sur 5 ans. Cioni *et. al* (2021, 24)

La crise de la reproductibilité

- Biais affectant le processus de publication :
  - Manipulation des valeurs p.
  - Faible puissance statistique.
  - Biais de confirmation des auteurs et relecteurs.

$\Longrightarrow$ Résultats publiés = faux positifs ? Ioannidis (2005)

La crise de la reproductibilité

- Biais affectant le processus de publication :
  - Manipulation des valeurs p.
  - Faible puissance statistique.
  - Biais de confirmation des auteurs et relecteurs.

  $\Longrightarrow$ Résultats publiés = faux positifs ? Ioannidis (2005)

- Quelques réponses :
  - Plans de gestion de données.
  - Plans de pré-analyse.
  - Méta-analyses.

La crise de la reproductibilité

- Biais affectant le processus de publication :
  - Manipulation des valeurs p.
  - Faible puissance statistique.
  - Biais de confirmation des auteurs et relecteurs.

  $\Longrightarrow$ Résultats publiés = faux positifs ? Ioannidis (2005)

- Quelques réponses :
  - Plans de gestion de données.
  - Plans de pré-analyse.
  - Méta-analyses.

- Condition nécessaire = capacité à reproduire les résultats publiés.

La crise de la reproductibilité

- Pas le cas par exemple en macro-économie

Résultats de reproduction de 67 articles en macro-économie. Chang et Li (2022, 11)

Les principes FAIR

- Les principes FAIR pour la reproductibilité (Wilkinson et al. 2016) :
  - Données trouvables (*Findable*).
  - Données accessibles (*Accessible*).
  - Données interopérables (*Interoperable*).
  - Données réutilisables (*Reusable*).
- Au cœur de la politique nationale de la recherche française :
  - Plan national pour la science ouverte.
  - Financement de la recherche par l'ANR.

Principes FAIR et incitations

- Poids de la documentation et mise en conformité FAIR = chercheurs.
- Contradiction avec incitations pour les chercheurs.
- Production de données peu valorisée :
  - Par les comités d'évaluation pour promotion.
  - Par les pairs, qui citent peu les données réutilisées :
    - SHS 2013 : 18 %. Robinson-García et al. (2015)
    - Biodiversité 2019 : 27 %. Kahn et al. (2021)

Principes FAIR et incitations

- L'outil *data paper* = outil d'alignement incitations :
  - Publication dédiée à des données = citation facilitée.
  - Publication dans une revue à comité de lecture = valorisation par comités d'évaluation.

Principes FAIR et incitations

- L'outil *data paper* = outil d'alignement incitations :
  - Publication dédiée à des données = citation facilitée.
  - Publication dans une revue à comité de lecture = valorisation par comités d'évaluation.
- Un cercle vertueux :
  - Meilleure documentation = utilisation adéquate et facilitée.
  - Conformité FAIR = données plus accessibles et diffusées.
  - Crible du comité de lecture = meilleure documentation et conformité FAIR.

Principes FAIR et incitations

- L'outil *data paper* = outil d'alignement incitations :
  - Publication dédiée à des données = citation facilitée.
  - Publication dans une revue à comité de lecture = valorisation par comités d'évaluation.
- Un cercle vertueux :
  - Meilleure documentation = utilisation adéquate et facilitée.
  - Conformité FAIR = données plus accessibles et diffusées.
  - Crible du comité de lecture = meilleure documentation et conformité FAIR.
- Le *data paper* contribue à résoudre le problème du passager clandestin des biens publics purs : la donnée documentée et FAIR.

L'écosystème TRF-GIS et reproductibilité

- Une réponse à la crise de reproductibilité en sciences (sociales). Wilkinson et al. (2016) Christensen et Miguel (2018)
- La Troisième République :
  - Production sans précédent de statistiques administratives. Desrosières (1993)
  - Administrations à différents niveaux d'agrégation.
  - Contexte de profonds changements socio-économiques.
- Renouveau de la recherche quantitative en histoire. Lemercier et Zalc (2019) Ruggles (2021)

L'écosystème TRF-GIS et reproductibilité

- Retard France vs autres pays (NHGIS, CAMPOP).
  - Quelques shapefiles départements et cantons, mais peu FAIR. LARHRA (2011)
  - Pas de cadre commun (nomenclatures), pas de documentation.
  - Perte de temps, peu d'interopérabilité, peu de crédibilité.

L'écosystème TRF-GIS et reproductibilité

- Quelques exceptions :
  - Cassini.ehess pour circonscriptions générales.
    Motte, Séguy, Théré (2003) Motte et Vouloir (2007)
  - HGIS infrastructures de transport.
    Thévenin et al. (2013) Mimeur et al. (2018)
  - HGIS communes (ANR-COMMUNES).
    Litvine, Séguy, Thévenin, Mimeur (2022)

$\Longrightarrow$ Le TRF-GIS établit un cadre commun documenté et reproductible.

Un *data paper* en SHS : pour qui ?

Où publier son data papers en SHS ?

- Une audience large et interdisciplinaire car peu de revues.
- *Data journals* :
  - En SHS :
    - *Journal of Open Humanities Data*.
    - *Research Data Journal for the Humanities and Social Sciences*.
  - Toutes disciplines :
    - *Data in Brief*.
    - *F1000 Research*.

Où publier son data papers en SHS ?

- Une audience large et interdisciplinaire car peu de revues.
- Revues acceptant explicitement des *data papers* en SHS :
  - *Cybergeo* pour la géographie.
  - *Historical Methods : A Journal of Quantitative and Interdisciplinary History* pour l'histoire.
  - *Frontiers in Sociology* pour la sociologie.
- D'autres revues peuvent accepter des *data papers*, e.g., en histoire :
  - *Explorations in Economic History*.
  - *Histoire & mesure*.

Audience des data papers

- Adaptation nécessaire de l'écriture :
  - Travail d'exposition du contexte en évitant le jardon disciplinaire.
  - Expliciter les présupposés de la collecte et des catégorisations.
  - Clarifier des usages potentiels.
  - Délimiter les contours du *data paper* (pas d'analyse ou généalogie).

Audience des data papers

- Adaptation nécessaire de l'écriture :
  - Travail d'exposition du contexte en évitant le jardon disciplinaire.
  - Expliciter les présupposés de la collecte et des catégorisations.
  - Clarifier des usages potentiels.
  - Délimiter les contours du *data paper* (pas d'analyse ou généalogie).
- Implications sur la forme des données :
  - Format le plus courant dans sa discipline (SPSS, Stata, SASS).
  - Format ouvert et universel (TXT, CSV).
- Implications sur l'entrepôt :
  - Français : Nakala, PROGEDO-ADISP.
  - International : Dataverse, Figshare, Zenodo.

Audience du data paper TRF-GIS

- Réutilisations possibles du TRF-GIS :
  - Cartographie de données historiques géolocalisées.
  - Appareillement de données pour analyse statistique.
- Audience : sciences sociales avec composante historique.
  - Histoire économique.
  - Démographie historique.
- Revue : *Historical Methods : A Journal of Quantitative and Interdisciplinary History*.
  - Large diffusion internationale.
  - Audience interdisciplinaire avec forte composante historique et dominante quantitative.

Diffusion des données TRF-GIS

- Format des données :
  - Format disciplinaire (histoire économique) : Stata *data format* (DTA).
  - Format ouvert et universel : TXT.
- Diffusion des données :
  - Harvard Dataverse : TRF-GIS Dataverse.
  - Data.gouv.fr : jeux de données TRF-GIS.
  - PROGEDO-ADISP : données historiques (bientôt).

Un *data paper* en SHS : comment ?

Un data paper en SHS : comment ?

- Des *data papers* générés par des machines ? Schöpfel et al. (2019)
  - Tendance à la génération automatique de *data paper* à partir des métadonnées.
  - Avantage : peu chronophage et diffusion rapide via moissonnage.
  - Mais ne permet pas la réutilisation adéquate car pas de description des méthodes de construction et choix de typologie.

$\Longrightarrow$ Travail minutieux de rédaction nécessaire (c.f. plus haut).

La structure d'un data paper

- Modèle *Scientific Data* du groupe Nature :

La structure d'un data paper

- Modèle *Scientific Data* du groupe Nature :

  1. Contexte et résumé : décrire succinctement les données, leur contexte et réutilisations potentielles.

La structure d'un data paper

- Modèle *Scientific Data* du groupe Nature :

  1. Contexte et résumé : décrire succinctement les données, leur contexte et réutilisations potentielles.
  2. Méthodes : décrire avec précision le processus de production et les choix de catégorisation.

La structure d'un data paper

- Modèle *Scientific Data* du groupe Nature :

  1. Contexte et résumé : décrire succinctement les données, leur contexte et réutilisations potentielles.
  2. Méthodes : décrire avec précision le processus de production et les choix de catégorisation.
  3. Fichiers de données : décrire chaque jeu de données (variables, noms de fichiers, localisation, formats, poids).

La structure d'un data paper

- Modèle *Scientific Data* du groupe Nature :

1. Contexte et résumé : décrire succinctement les données, leur contexte et réutilisations potentielles.
2. Méthodes : décrire avec précision le processus de production et les choix de catégorisation.
3. Fichiers de données : décrire chaque jeu de données (variables, noms de fichiers, localisation, formats, poids).
4. Validation technique : décrire les procédures confirmant la validité des données.

La structure d'un data paper

- Modèle *Scientific Data* du groupe Nature :

1. Contexte et résumé : décrire succinctement les données, leur contexte et réutilisations potentielles.
2. Méthodes : décrire avec précision le processus de production et les choix de catégorisation.
3. Fichiers de données : décrire chaque jeu de données (variables, noms de fichiers, localisation, formats, poids).
4. Validation technique : décrire les procédures confirmant la validité des données.
5. Notes d'usage : décrire les procédures concrètes de réutilisation des données.

La structure d'un data paper

- Modèle *Scientific Data* du groupe Nature :

1. Contexte et résumé : décrire succinctement les données, leur contexte et réutilisations potentielles.
2. Méthodes : décrire avec précision le processus de production et les choix de catégorisation.
3. Fichiers de données : décrire chaque jeu de données (variables, noms de fichiers, localisation, formats, poids).
4. Validation technique : décrire les procédures confirmant la validité des données.
5. Notes d'usage : décrire les procédures concrètes de réutilisation des données.
6. Disponibilité du code : décrire les procédures d'accès au code de reproduction.

La lien données-article

- Un écosystème données-article :
  - Accès aux données dès la soumission de l'article.
  - Concrètement : identifiant pérenne dans le texte (DOI).
  - Possibilités d'attribuer un DOI accessible anonymement par un code pour la relecture (dépend de l'entrepôt).

Le data paper « Mapping the Third Republic »

- Suit le modèle de *Scientific Data*.
- Méthodes :
  - Décrit la méthodologie technique de production et ses limites.
  - Détaille les éléments institutionnels relatifs aux administrations, e.g., :
    - Perte et retour de l'Alsace-Lorraine.
    - Réforme des arrondissements de 1926.
    - Réformes militaires de 1873-1874.
    - Lois de redécoupage électoral.
- Validation technique :
  - Confrontation des sources secondaires à 175 sources primaires.
  - Confrontation au *shapefile* des cantons de 1884 du LARHRA.

Conclusion

- Apprentissage nécessaire pour ce nouveau type d'article en SHS.
- Nombreuses ressources à disposition des chercheurs :
  - Séminaires, journées spéciales, colloques...
  - Dossiers d'acteurs institutionnels : DoRANum et CoopIST (Cirad).
  - Offres de formation : URFIST, INRAE, PUD / PROGEDO.

Merci !

- TRF-GIS Dataverse
  https://dataverse.harvard.edu/dataverse/TRF-GIS

- Gay, Victor. 2021. « Mapping the Third Republic. A Geographic Information System of France (1870-1940) ». *Historical Methods : A Journal of Quantitative and Interdisciplinary History*. À paraître.

  https://doi.org/10.1080/01615440.2021.1937421
  https://hal.archives-ouvertes.fr/hal-02951461

- Contact :

  - Email : victor.gay@tse-fr.eu
  - Page personnelle : https://www.victorgay.me
  - Twitter : @victorgayeco