Les enjeux et la mise en place de liens entre les produits de recherche

Maud Medves (INRIA)
Alexis Arnaud (UGA)
Maria Grazia Santangelo (UGA)
Elias Chetouane (UGA)

Open Science Days - 2023-11-16

Plan

Plan

État des lieux

Increasing the positive impacts of journal policies on data sharing practices, Breakout 4, Research Data Alliance Plenary 20, Gothenburg, 2023

- 2016-2019 : augmentation du nombre de déclaration de disponibilité des données
- Modes principaux de diffusion des données :
  - 37% → 35% : dans un entrepôt de données
  - 23% → 15% : non indiqué
  - 10% → 16% : sur demande

- Increasing the positive impacts of journal policies on data sharing practices (video).
- Hrynaszkiewicz, I., Simons, N., Hussain, A., Grant, R. and Goudie, S., 2020. Developing a Research Data Policy Framework for All Journals and Publishers. Data Science Journal, 19(1), p.5. http://doi.org/10.5334/dsj-2020-005.

État des lieux

Data sharing practices and data Analysis availability upon request differ across scientific disciplines

- Étude de la disponibilité des "données critiques" (jeu de données, image, modèle) :
  - dans 875 articles de 9 disciplines
  - publiés dans Nature and Science
  - sur les périodes 2000-2009 et 2010-2019

- Disponibilité initiale :
  - complète 54.2% (33.0 – 82.8%)
  - partielle 71.8% (40.4 – 100%)
  - différences significatives : par domaine de recherche, type de données, journal, période de publication

- Tedersoo, L., Küngas, R., Oras, E. et al. Data sharing practices and data availability upon request differ across scientific disciplines. Sci Data 8, 192 (2021). https://doi.org/10.1038/s41597-021-00981-0

État des lieux

Data sharing practices and data Analysis availability upon request differ across scientific disciplines

- 310 auteurs ont été contactés :
  - 39.4% (27.9 – 56.1%) de réponses positives
  - 19.4% de réponses négatives : fortes différences par domaine
  - délai médian de 15 jours pour obtenir les données
  - augmentation de la disponibilité globale des données après un délai de 60 jours :
    - ⋆ complète 69.5% (57.0 – 87.9%)
    - ⋆ partielle 83.2% (64.9 – 100%)
    - ⋆ plus faible disponibilité pour la période 2000-2009 (29.4% vs 56.0%)
- Lieux de diffusion des données :
  - 62.2% : supplementary materials
  - 35.0% : pas de lieu indiqué
  - 22.3% : archives de données
  - 19.7% : sur demande

  - certains jeux de données sont déposés dans plusieurs lieux

État des lieux

A study of the impact of data sharing on article citations using journal policies as a natural experiment

- 2 études sur l'impact de la diffusion des données sur le nombre de citations :
  - 17 journaux avec une politique de partage des données, 13 sans politique de partage
  - 200 articles (ou 2 années de publication) par journal
  - utilisation de Web of Science pour le suivi des citation sur les 5 ans après publication
  - pour la 2e étude : 2 journaux avec et 2 journaux sans politique forte de diffusion des données
- Résultats
  - étude 1 :
    - pas de différence significative dans le nombre de citations entre avant et après la mise en place d'une politique de diffusion des données
  - étude 2 :
    - 25 $\rightarrow$ 70% pour la diffusion des données entre avant et après le changement de politique
    - 40% d'augmentation des citations après le passage à une politique forte

- Christensen G, Dafoe A, Miguel E, Moore DA, Rose AK (2019) A study of the impact of data sharing on article citations using journal policies as a natural experiment. PLOS ONE 14(12) : e0225883. https://doi.org/10.1371/journal.pone.0225883

Enjeux de lier les produits de recherche entre eux

- Pour soi :
  - Augmentation de la visibilité (et donc des citations) de ses publications.
  - Valorisation de tous les produits de recherche.
    - ⋆ OSD@UGA 2022 : demi-journée consacrée à la prise en compte du logiciel dans l'évaluation des carrières en recherche.
- Pour les autres :
  - Nécessité pour la reproductibilité et la transparence.
  - Permet de découvrir des produits de recherche de proche en proche.
  - Valorisation de tous les membres d'une équipe de recherche.

- En prime : l'ANR et les financements européens imposent l'accès libre des publications et des données, de ce fait l'effort de diffusion étant fait, il est simple de faire des liens entre produits de recherche.

Plan

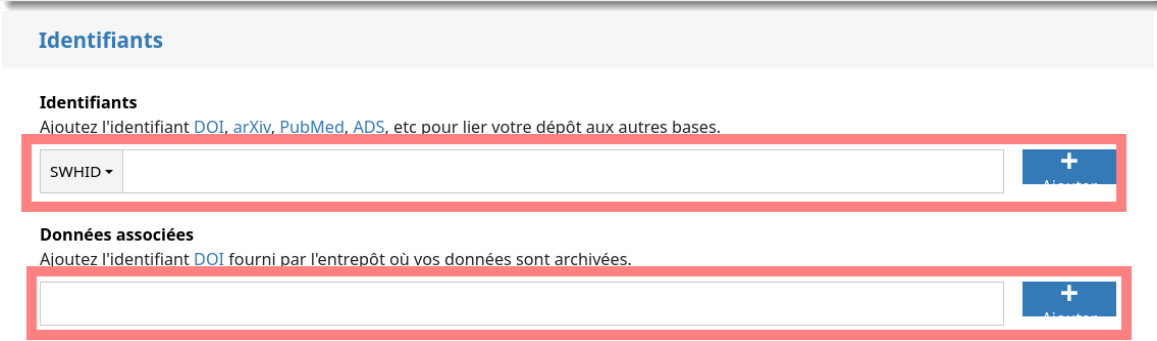
Lier les publications, les données et les codes

Du côté des publications

Depuis HAL

Remplir les métadonnées **“Identifiants”** :

- codes : SWHID
- données : DOI



Lier les publications, les données et les codes

Du côté des données

Depuis un entrepôt - ex. : Recherche Data Gouv

Remplir les métadonnées :

- **“Publications associée”** : HAL-ID ou DOI

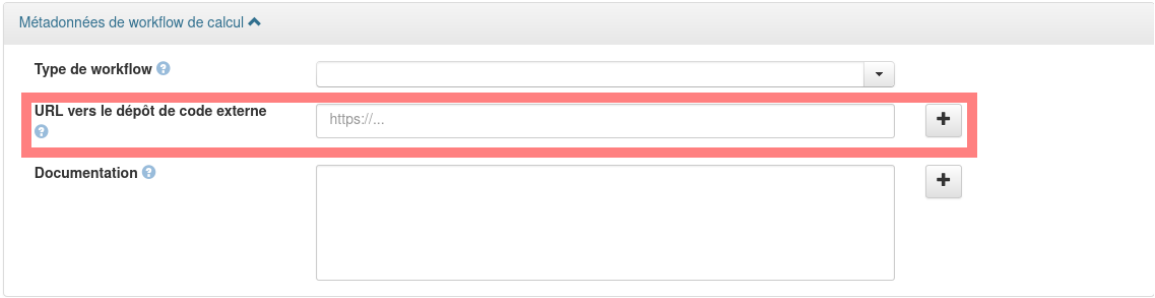
Lier les publications, les données et les codes

Du côté des données

Depuis un entrepôt - ex. : Recherche Data Gouv

Remplir les métadonnées :

- **“Workflow de calcul”** : SWHID ou url vers Software Heritage



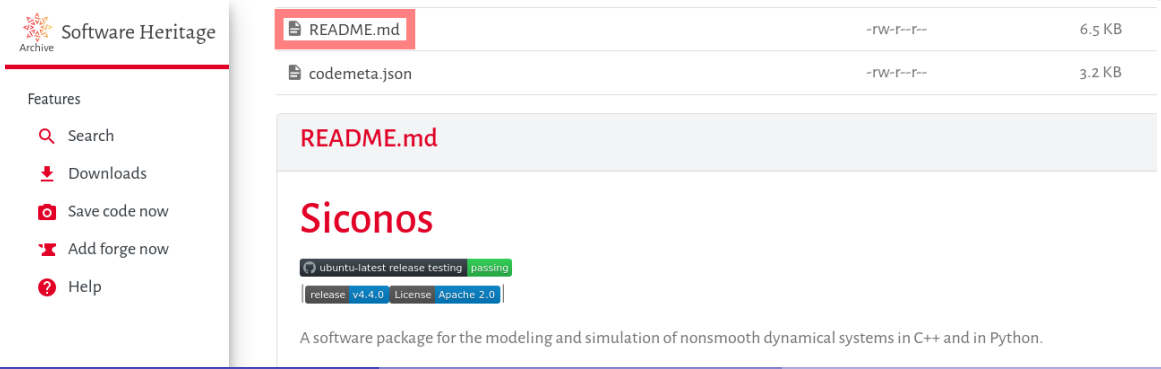
Lier les publications, les données et les codes

Du côté des codes

Depuis Software Heritage

Remplir le fichier **“README”** :

- publications : HAL-ID ou DOI
- données : DOI



Plan

Projet FAIRCORE4EOSC

Cas concret de lien entre publications, données et codes

- Projet Horizon Europe financé sur la période juin 2022 à mai 2025
- En phase de spécification technique
- La phase de beta release démarrera fin novembre 2023

Projet FAIRCORE4EOSC

Présentation du projet

- 23 partenaires dont Inria via Software Heritage
- Objectifs : développer et mettre en oeuvre neuf composants de base de l'EOSC. Pour améliorer la découvrabilité et l'interopérabilité d'un nombre toujours croissant de produits de la recherche.
- Les composants de base : 3 exemples
  - MSCR : Metadata Schema and Crosswalk registry
  - SWHM : Miroir de l'archive Software Heritage
  - RSAC : Research Software APIs and Connectors

Projet FAIRCORE4EOSC

Work package 6 - *Research Software APIs and Connectors* (RSAC)

- Le WP6 et le travail mené par les RSAC
- Qui sont les RSAC ?
  - Deux archives ouvertes : InvenioRDM au CERN, Dataverse à DANS,
  - deux éditeurs : Dagstuhl et Episciences,
  - deux aggrégateurs : swMATH et OpenAire.

Projet FAIRCORE4EOSC

RSAC et le rapport *Scholarly Infrastructures of Research Software* (SIRS)

- La base de travail : le rapport SIRS (scholarly infrastructures of research software) :
- Les quatre piliers : archiver le code source, le référencer, le décrire, le citer.
  - Archiver => s'assurer qu'on peut accéder au code source,
  - Référencer => s'assurer que l'on peut identifier les codes sources,
  - Décrire => faciliter la découverte du code source,
  - Citer => créditer les auteurs pour rendre le travail de développement attractif en matière de carrière.

Projet FAIRCORE4EOSC

Exemple de réalisation

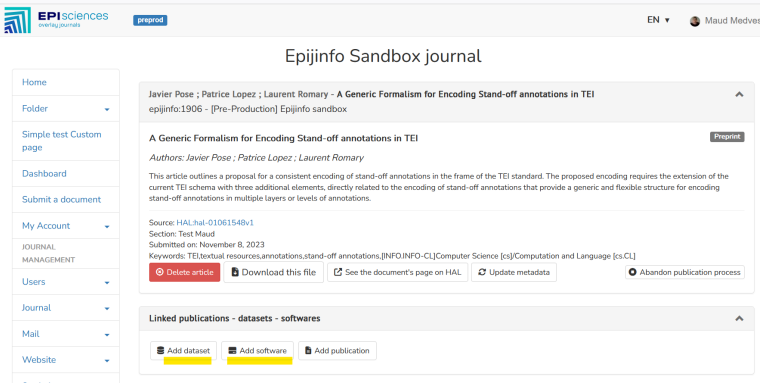
- concrètement, ça signifie quoi ?
- ex. d'une archive ouverte (Zenodo)
  - dépôt dans Software Heritage des codes sources déposés dans Zenodo ;
  - obtention du SWHID associé ;
  - exposition du SWHID dans la notice maintenue par l'archive ouverte
- ex. d'un aggrégateur (swMath)
  - extraire les références à des codes sources présentes dans des publications et déclencher l'archivage de ce code source connu de l'aggrégateur mais absent de Software Heritage ;
  - exposer le SWHID correspondant dans la notice du code source maintenue par l'aggrégateur

Episciences

Associer code source, données et manuscrit dès la soumission

- Développement de nouvelles fonctionnalités d'Episciences dans le cadre du projet FAIRCORE4EOSC (actuellement disponible en test)

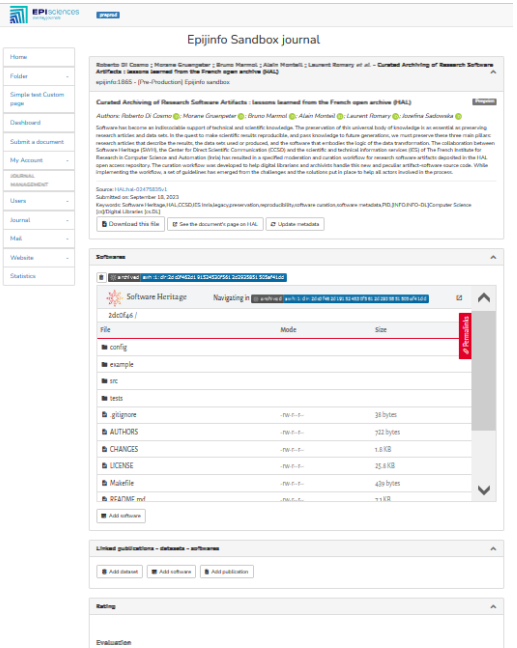
- Possibilité d'associer du code source ou des données au manuscrit que l'auteur soumet à une revue (via un identifiant, swhid ou DOI).



Episciences

Données et codes dans l'interface de relecture

- Le code / les données sont visibles par les relecteurs dans le formulaire de relecture.

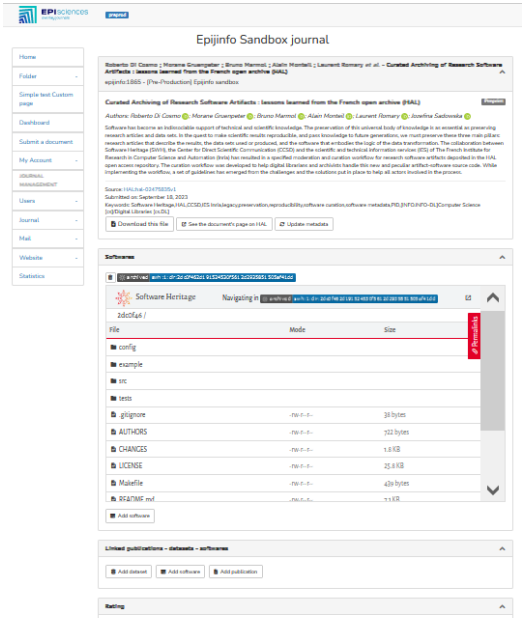


Episciences

Mettre en valeur données et codes associés sur la page de l'article

- Le code / les données sont affichés sur la page de lecture de l'article (et mentionnés dans le pdf)

- Un progrès important par rapport aux déclarations de disponibilité des données (DAS en anglais) actuelles.



Plan

En bref

- L'indication du lieu de diffusion des données dans les publications n'est pas assez présente.
- Ce lieu de diffusion n'est pas systématiquement pérenne.
- Le lien entre produits de recherche améliore la visibilité de tous les produits de recherche.
- Pour le moment, les liens sont à la charge des équipes de recherche, via HAL et les entrepôts de données.
- Mais une structuration forte en cours au niveau européen, notamment pour reporter la charge des liens au niveau des éditeurs ou diffuseurs.

Contacts

Merci de votre attention !

Site web de la science ouverte à l'UGA.