



**HAL**  
open science

## **BibHelioTech**

Vincent Génot, Axel Dablanc, Guillaume Cabanac, Camille de Salabert, Sabine Barreaux, Pascal Cuxac, Nicolas Dufourg, Nicolas Aunai, Williams Exbrayat

► **To cite this version:**

Vincent Génot, Axel Dablanc, Guillaume Cabanac, Camille de Salabert, Sabine Barreaux, et al.. BibHelioTech. Colloque du Programme National Soleil-Terre (PNST 2022), May 2022, Marseille, France. . hal-04284893

**HAL Id: hal-04284893**

**<https://hal.science/hal-04284893>**

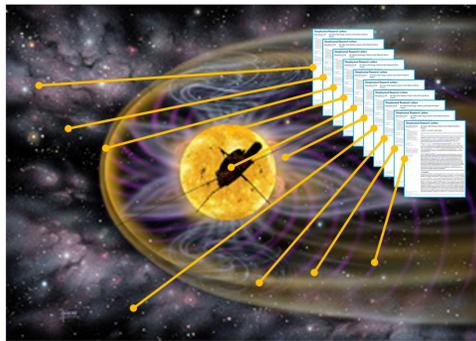
Submitted on 14 Nov 2023

**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.

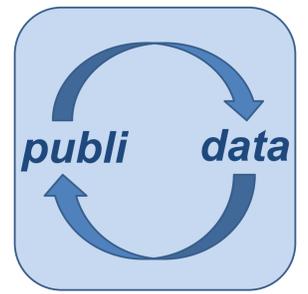
L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

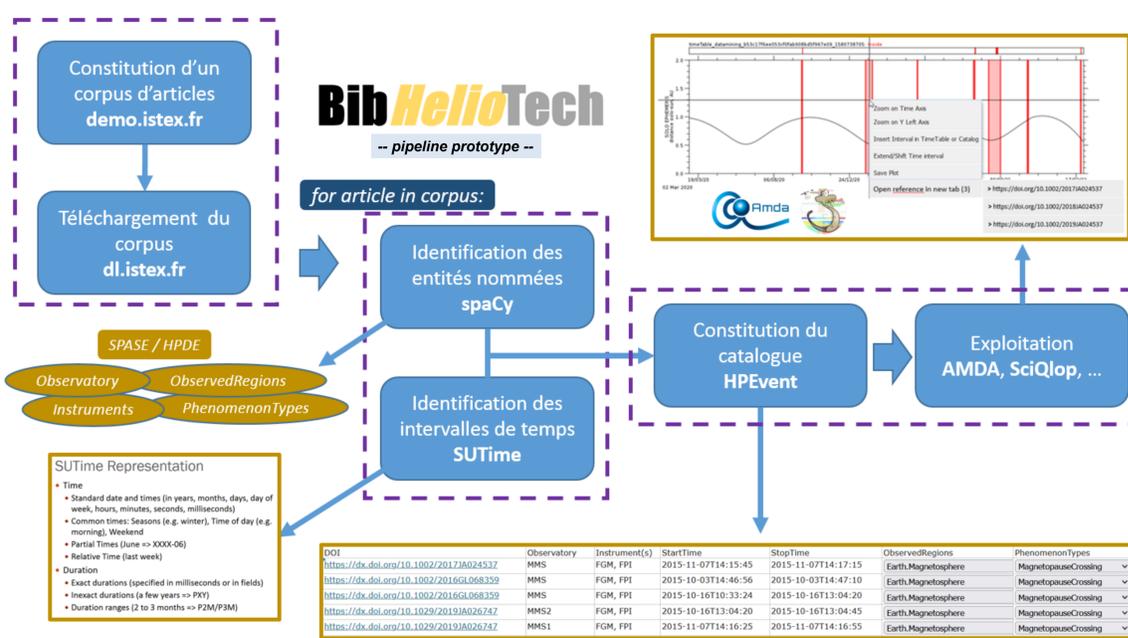
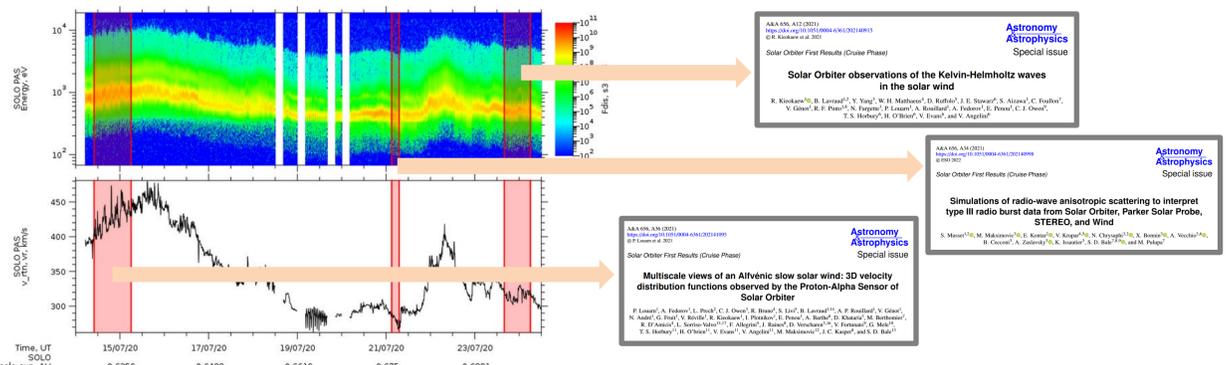


A partir d'un corpus d'articles scientifiques du domaine héliophysique utilisant des données de missions spatiales, nous réaliserons une détection textuelle automatisée sur les événements observés, les satellites/instruments utilisés, les régions spatiales et les processus physiques concernés, afin de relier ces entités avec les publications dont elles sont extraites, dans des catalogues exploitables par les outils d'analyse de données de la discipline. Ce lien fort et systématisé entre données et publications, inexistant à ce jour, augmentera l'expérience d'analyse de données en immergeant le chercheur dans le contexte bibliographique de son cas d'étude, améliorera significativement la reproductibilité des résultats publiés, et facilitera la réutilisation de ces catalogues dans de nouvelles études statistiques et comparatives.



**Cas d'utilisation**

- Visualisation de données Solar Orbiter PAS dans AMDA en Juillet 2020. Les 3 intervalles en rouge correspondent à ceux étudiés dans les 3 articles à droite.
- Le but de BibHelioTech est de produire des catalogues, disponibles dans les outils, liant publications et intervalles d'étude par les missions/instruments.

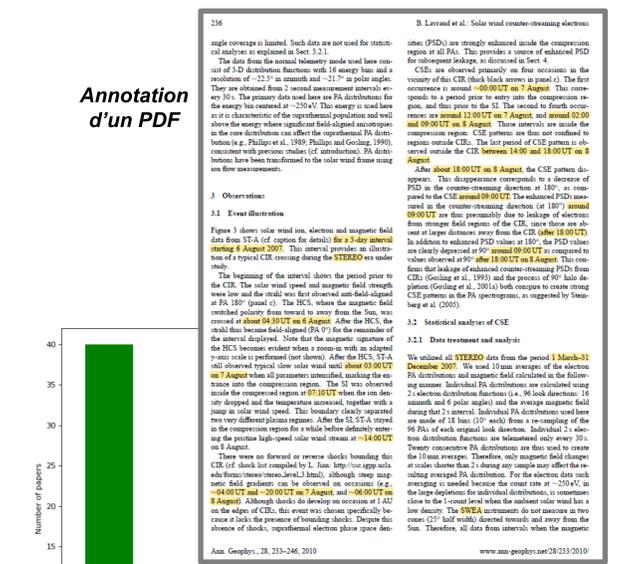


The screenshot shows the ISTEX website, described as 'Le plus vaste réservoir d'archives scientifiques au service de la recherche française'. It features statistics: 25,6 M de documents, 9 318 revues, 349 009 ebooks, and 1473 publications in 2020. A search bar is visible with a query: 'https://api.istex.fr/document/?q=magnetopause&face=corpusName[1]&size=20&sortBy=publicationDate[desc]&output=&tats'. Below the search bar, a list of search results is shown, including a paper titled 'Effect of the solar wind density on the evolution of normal and inverse cor...'.

## Etapes / avancement du projet (1.5 mois du stage d'A. Dabianc)

- Annotation de +50 articles en se focalisant sur intervalles de temps, missions et instruments (tels que dispo sur le registre SPASE <https://hpde.io/SMWG/index.html>)
  - Récupération automatique des annotations (bibliothèque PyMuPDF)
- Océrisation = PDF → image → TXT :
  - TESSERACT (bibliothèque python qui convertit le PDF en texte brut)
    - Lourd, mais toutes les infos sont conservées
  - GROBID fournit un fichier XML structuré
    - Rate beaucoup de données, dont les images et légendes
- Reconnaissance automatique des DOI, 2 approches complémentaires efficaces :
  - GROBID à partir de la balise DOI
  - Si DOI absent de GROBID : à partir du titre de l'article et des API de NASA / ADS
- SUTime : reconnaissance d'entités temporelles, utilisé pour les intervalles de temps
  - Performant (balise « duration »), mais pas encore optimisé
  - Métriques obtenues à partir d'une comparaison avec les intervalles annotés
- NER (Named Entity Recognition), 2 approches :
  - utilisation de FLAIR (bibliothèque python)
    - 1<sup>er</sup> essai d'entraînement d'un modèle avec de nouvelles entités INST, SAT, REG, PROCS
    - Performance (entraînement sans contexte) : voir métriques →
    - Comparaison textuelle (plus légère en temps de traitement/CPU)
- TODO : association entre les intervalles et missions/instruments détectés
  - C'est une des difficultés identifiées !
  - Nous testerons une approche statistique en fonction de l'occurrence entre éléments et de leur proximité dans l'article

## Annotation d'un PDF



**Résultat de l'utilisation de la bibliothèque FLAIR pour l'entraînement d'un modèle avec les nouvelles classes INST, SAT, REG, PROCS issues du modèle de données SPASE**

2022-05-02 13:53:58,397 0.8092 0.8113 0.8103 0.8038  
 2022-05-02 13:53:58,358

Results:

- F-score (micro) 0.8103
- F-score (macro) 0.7393
- Accuracy 0.8038

By class:

	precision	recall	F1-score	support
INST	0.8060	0.9562	0.8739	35343
SAT	0.8185	0.9592	0.8892	19404
REG	0.9983	0.5882	0.7403	3927
LOC	0.8100	0.9816	0.8934	1646
ORG	0.7458	0.7528	0.7493	1715
PER	0.9387	0.9277	0.9332	1618
PROCS	0.2500	0.5000	0.3333	462
MISC	0.7864	0.7026	0.7421	723
micro avg	0.8092	0.8113	0.8103	64838
macro avg	0.7692	0.7403	0.7393	64838
weighted avg	0.8190	0.8113	0.8029	64838