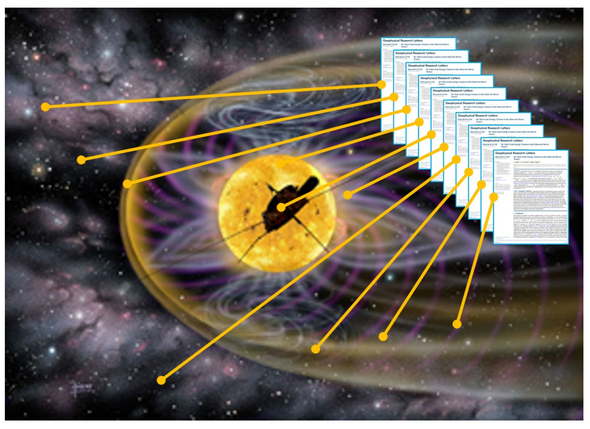
BibHelioTech

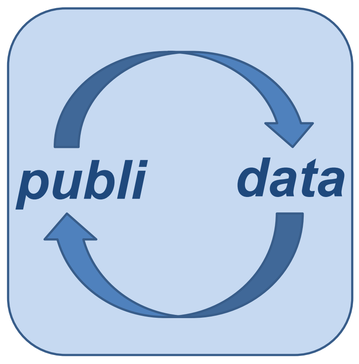
Vincent Génot, Axel Dablanc, Guillaume Cabanac, Camille de Salabert, Sabine Barreaux, Pascal Cuxac, Nicolas Dufourg, Nicolas Aunai, Williams Exbrayat

BibHelioTech

Vincent Génot, Axel Dablanc, Guillaume Cabanac, INSU/IRAP et IRIT (Toulouse)
Camille de Salabert, Sabine Barreaux, Pascal Cuxac, Inist-CNRS (Nancy)
Nicolas Dufourg, CNES/DOA (Toulouse)
Nicolas Aunai, INSIS/INSU/LPP (Palaiseau)
Williams Exbrayat, INSU/Observatoire Midi-Pyrénées/UAR831 (Toulouse)

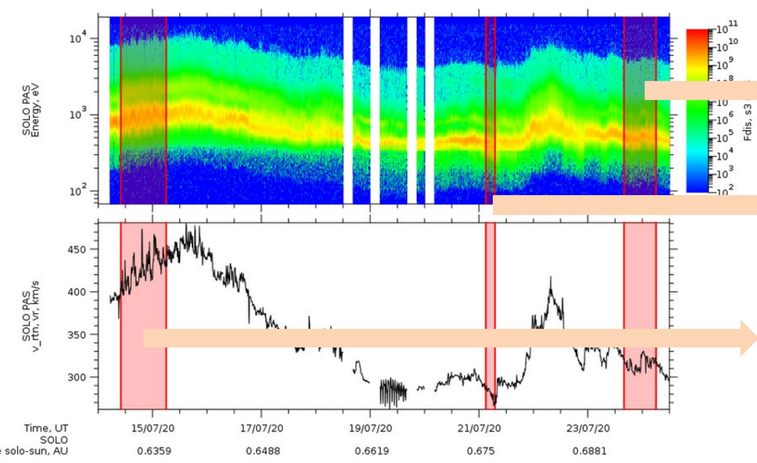


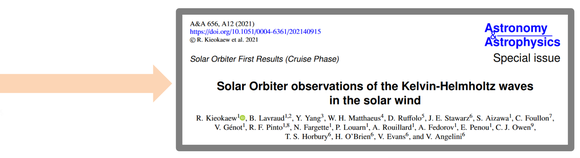
A partir d'un corpus d'articles scientifiques du domaine héliophysique utilisant des données de missions spatiales, nous réaliserons une détection textuelle automatisée sur les événements observés, les satellites/instruments utilisés, les régions spatiales et les processus physiques concernés, afin de 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.

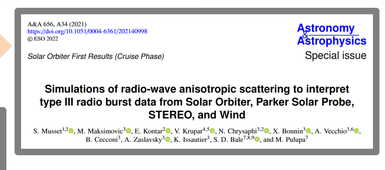


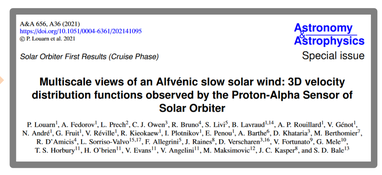
Cas d'utilisation

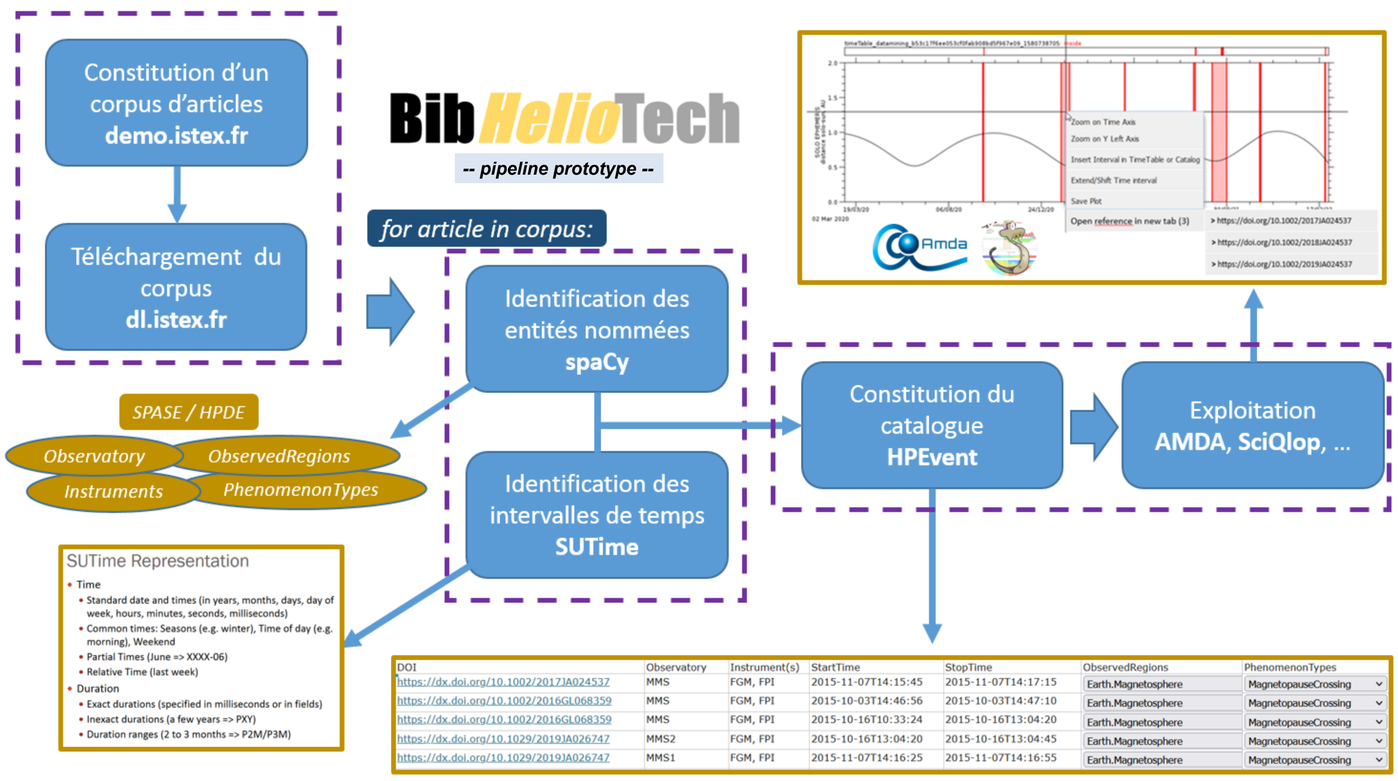
- Visualisation de données Solar Orbiter PAS dans **AMDA** en Juillet 2020. Les 3 intervalles en rouge correspondent à ceux étudiés dans les 3 articles à droite.
- Le but de **Bib*Helio*Tech** est de produire des catalogues, disponibles dans les outils, liant publications et intervalles d'étude par les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









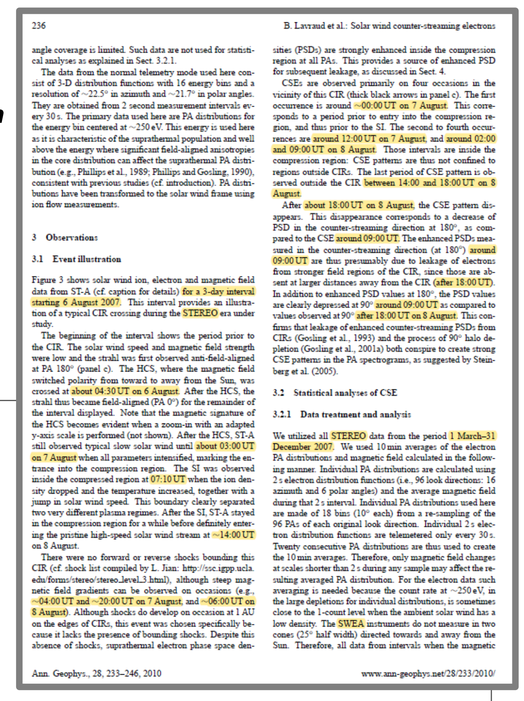


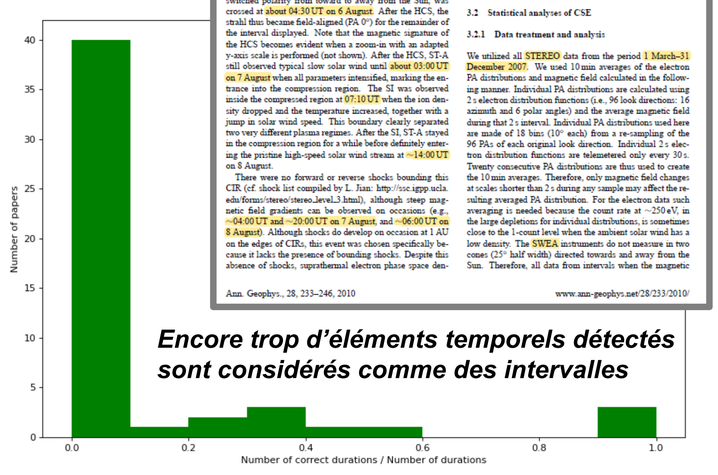
DOI	Observatory	Instrument(s)	StartTime	StopTime	ObservedRegions	PhenomenonTypes
https://dx.doi.org/10.1002/2017JA024537	MMS	FGM, FPI	2015-11-07T14:15:45	2015-11-07T14:17:15	Earth.Magnetosphere	MagnetopauseCrossing
https://dx.doi.org/10.1002/2016GL068359	MMS	FGM, FPI	2015-10-03T14:46:56	2015-10-03T14:47:10	Earth.Magnetosphere	MagnetopauseCrossing
https://dx.doi.org/10.1002/2016GL068359	MMS	FGM, FPI	2015-10-16T10:33:24	2015-10-16T13:04:20	Earth.Magnetosphere	MagnetopauseCrossing
https://dx.doi.org/10.1029/2019JA026747	MMS2	FGM, FPI	2015-10-16T13:04:20	2015-10-16T13:04:45	Earth.Magnetosphere	MagnetopauseCrossing
https://dx.doi.org/10.1029/2019JA026747	MMS1	FGM, FPI	2015-11-07T14:16:25	2015-11-07T14:16:55	Earth.Magnetosphere	MagnetopauseCrossing

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

- 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 (librairie **PyMuPDF**)
- **Océrisation** = PDF → image → TXT :
  - **TESSERACT** (librairie 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** (librairie python)
    - 1er essai d'entrainement d'un modèle avec de nouvelles entités ***INST, SAT, REG, PROCS***
    - Performance (entrainement 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





Encore trop d'éléments temporels détectés sont considérés comme des intervalles

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

```
2022-05-02 13:53:58,357 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    0.9542    0.8739     35343
         SAT     0.8185    0.5952    0.6892     19404
         REG     0.9983    0.5882    0.7403      3927
         LOC     0.8100    0.9016    0.8534      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
```

Colloque du Programme National Soleil Terre, Marseille, 16-20 mai 2022