Convention de structure des données de référence au format JSON

Dans un fichier .json de (méta)données de référence, les données sont organisées comme suit :

- À la plus grande échelle, le fichier se présente une liste de dictionnaires où chaque dictionnaire, représenté par {…} sur le patron, correspond aux données d'un article de recherche.

[{…}, {…}, … , {…}]

- Chaque dictionnaire dans cette liste se présente sous la forme suivante :

```
{
"file_path": "/…/…/nom_du_fichier.pdf",
"file_url": "https://tel.archives-ouvertes.fr/tel-01653192/document",
"title": "Évolution des Architectures des Systèmes Avioniques Embarqués",
"year": "2016",
"authors": {…},
"global_Structure_to_Parent_MAP": { … } ,
"unassigned_affiliations": [ … ],
"longString" : "Abstract : …."
}
```

Les champs en vert sont toujours nécessairement présents dans une donnée valide. Les deux champs en bleu ne sont pas forcément simultanément présents, mais ne peuvent être simultanément absents. Les champs en rouge sont des données supplémentaires qui ne sont pas forcément présentes.

Bien sûr, des données à usage spécifique pourraient accepter d'avoir certains champs manquants, par exemple l'année n'est pas nécessaire si on cherche à établir des données de référence pour évaluer notre capacité à trouver les bonnes affiliations. Mais on ne considère une donnée comme valide seulement si elle vérifie les conditions précédentes.

- Le champ "authors" à pour valeur un dictionnaire qui associe à des noms d'auteurs une liste d'affiliations, de la forme suivante :

```
{
"Boris Moret":
[
"Institut Polytechnique de Bordeaux",
"Université Sciences et Technologies - Bordeaux 1",
"Laboratoire de l'intégration, du matériau au système",
"Centre National de la Recherche Scientifique"
]
}
```

Le(s) prénom(s) (potentiellement absents) viennent avant le nom d'auteur.

- Le champ "global_Structure_to_Parent_MAP") pour valeur un dictionnaire associant à chaque structure ses structures parentes. Cela permet d'organiser les affiliations d'un auteur en plusieurs arbres qui représentent « l'emboîtement » des affiliations.

Le champ "unassigned_affiliations" peut contenir une liste d'affiliations qui n'étaient pas assignés à un auteur, ou assignés à un auteur dont on ne connaît pas le nom.

Le champ "longString" contient le texte récupéré du pdf à partir duquel on va chercher les affiliations et autres.