GROBID ou comment ouvrir les portes des traitements analytiques aux archives ouvertes et aux bases de publications

Patrice Lopez – SCIENCE-MINER[1]

Résumé

Dans leur grande majorité, les publications académiques ne sont disponibles qu'en PDF, format très peu adapté aux analyses de corpus. Dans le but d'ouvrir les archives ouvertes à de nouvelles applications et à des analyses bibliométriques, le développement de GROBID a démarré en 2008, pour aboutir aujourd'hui à un outil de référence en extraction d'information bibliographiques structurées, utilisé dans de nombreuses archives ouvertes et services IST commerciaux.

Au delà des informations bibliographiques, nous présentons de nouvelles structurations automatiques réalisées par GROBID dans le but de rendre possible de nouvelles analyses de corpus : 1) l'ingestion complète des PDF dans un format TEI riche et uniforme, 2) l'extraction des figures et tables, 3) la génération de PDF enrichis par des annotations structurelles et 4) l'extraction des mesures physiques. Ces différents exemples illustrent les contraintes pratiques de l'exploitation des bases de publication et la nécessité, présente et future, d'outils tels GROBID pour accompagner cette ambition.

La difficile mise en pratique des méthodes analytiques sur les corpus de publications

Outre les problèmes bien connus de couverture et de droits d'utilisation, de nombreuses méthodes d'analyse de corpus de publications académiques présupposent la disponibilité de pleins textes d'une qualité quasi-parfaite, voir considèrent l'existence d'un format de document structuré XML comme point de départ. Cependant la grande majorité des corpus de publications est au format PDF. Pour ISTEX, où pourtant les XML pleins textes éditeurs sont très largement surreprésentés, plus de 91% des articles ne sont disponibles qu'en PDF.

Cela signifie qu'en pratique, le texte disponible est fortement bruité et que l'ensemble des structures, même les plus élémentaires (résumés, paragraphes, mots clefs, etc.), sont absentes pour tout traitement informatique. Cela implique encore que la plupart des techniques actuelles d'analyse de corpus ne sont, au mieux, que partiellement applicables sur la base de métadonnées éditeurs très réduites, voir inapplicables - c'est le cas, par exemple, des études d'impact qui dépendent de la disponibilité à très grande échelle de références bibliographiques et des affiliations détaillées.

D'autre part, même pour les meilleurs formats XML éditeurs actuels, le niveau de structuration est souvent partiel (ex. PubMedCentral) et s'avère extrêmement hétérogène d'un éditeur à l'autre. Cette hétérogénéité impose à la fois un travail considérable d'uniformisation des formats, mais également un effet "plus petit dénominateur commun" où le niveau de structuration commun possible est de facto le plus faible de l'ensemble d'un corpus.

GROBID

L'objectif de GROBID [1] est d'ingérer des documents au format PDF - ou de meilleure qualité si disponibles - afin de produire automatiquement à la fois :

- un format structuré uniforme et sémantiquement riche basé sur la TEI [2], adapté à des traitements analytiques, indexations, créations de graphe, etc. à grande échelle,
- des PDF enrichis apportant aux utilisateurs une valeur ajoutée facilement appréhendable.

Initié en octobre 2008, GROBID est un logiciel Open Source depuis février 2011, reposant aujourd'hui entièrement sur de l'apprentissage automatique (linear CRFs). GROBID correspond à l'état de l'art actuel en extraction et structuration automatique des metadonnées d'en-tête (titre, auteurs,

[1] Pour information, l'auteur est également collaborateur externe de l'équipe Alpage Inria et consultant technique pour le projet ISTEX.

affiliations, etc., voir [2]) et des références bibliographiques (sélection compétitive pour les projets Semantic Scholar et ISTEX). Outre la qualité des extractions, GROBID est robuste, rapide, compact en usage mémoire, et peut monter en charge sans difficulté en mode distribué sur plusieurs millions de documents.

GROBID est ainsi déployé chez d'importants utilisateurs de bases de publications et des archives ouvertes telles le CERN, le JPL (NASA), l'INIST (projet ISTEX), le CCSD (HAL), l'Office Européen des Brevets, ou encore l'Allen Institute for Artificial Intelligence (Semantic Scholar [3]), mais également commerciaux comme ResearchGate ou Mendeley (Elsevier).

Plus de structures fiables pour plus d'analyses

Cette présentation abordera plus particulièrement de nouvelles extractions supportées depuis peu par GROBID, rendant possible de nouvelles applications pratiques sur des bases de publications. Nous présentons à la fois ces nouveaux résultats et les types d'analyse rendus possibles qui nous semblent d'intérêt dans le cadre de cette journée d'étude.

- **une structuration TEI fiable des corps de texte**, c'est-à-dire offrant divisions en paragraphes, titres de section, notes, etc. incluant les marqueurs de références, de figures et de tables. Cette structure de corps de texte permet d'ancrer précisément des annotations textuelles automatiques ou manuelles, d'améliorer des techniques d'extraction et de fouille de textes, d'exploiter des contextes de citations, mais également d'attacher des informations d'interactions au niveau du texte (par exemple les extraits les plus souvent annotés ou soulignés).

- **l'extraction des figures et tables**, incluant ici titre de figure, légende, zone de la figure dans le PDF. De telles extractions présentent des intérêts applicatifs pour la présentation d'article (une des utilisations de GROBID à ResearchGate). Elle sont nécessaires à la recherche de figures ou encore l'analyse et la classification d'images d'un sous-corpus thématique de publications (telles les familles d'images astronomiques).

- **la production de PDF directement enrichis par des annotations structurelles**. Par exemple, nous produisons à l'aide de services OpenURL des PDF dont les références bibliographiques sont directement cliquables si elles appartiennent à un ensemble de ressources accessibles, renvoyant alors vers les PDF cités. L'intérêt est ici de pouvoir présenter des annotations immédiatement appréhendables par un utilisateur final, et non pas se limiter au stockage des annotations pour des traitements machine. D'autre part, cela rend possible l'ajout d'une couche interactive à un PDF (lien vers des articles Wikipedia, des pages d'auteurs, recommandations, etc.).

- **l'identification et la normalisation des mesures physiques**. L'objectif de cette structuration est de capturer les nombreuses mesures physiques présentes dans les textes scientifiques (mesures de pression, température, etc.) sous une forme exploitable par une machine, ceci malgré une perte de qualité due au PDF et à l'utilisation éventuel d'un OCR. De telles structurations et normalisations sont la condition obligatoire pour la recherche fiable d'expressions de quantité dans un corpus, mais également à plus long terme utiles à l'enrichissement de bases de connaissances et à la génération automatique d'hypothèses scientifiques.

Nous pensons que les différentes tâches accomplies par GROBID correspondent à des briques qui sont ou seront de plus en plus indispensables aux archives ouvertes et aux outils d'analyse des publications académiques. Dans le cas de GROBID, un travail en apprentissage automatique au niveau de l'état de l'art du domaine a été combiné à une réalisation Open Source et un important effort d'ingénierie. Ce chantier-triple vise à faciliter une utilisation performante, pratique et large, profitable à l'ensemble de la communauté IST et donc aux chercheurs. En retour, nous bénéficions de contributions et de cas d'usage qui permettent une amélioration continue de l'outil - outil qui lui-même rend possible nos propres travaux d'analyse de bases de publications.

Références

[1] GROBID (2008-2016) https://github.com/kermitt2/grobid, http://grobid.readthedocs.org

[2] TEI: Text Encoding Initiative (1990-2016) http://www.tei-c.org

[3] M. Lipinski, K. Yao, C. Breitinger, J. Beel, and B. Gipp, Evaluation of Header Metadata Extraction Approaches and Tools for Scientific PDF Documents, in Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL), Indianapolis, IN, USA, 2013.

[4] https://www.semanticscholar.org