

Projet d’extraction de données de type semi-structurées (tables)

au sein des documents PDF

Par  
Frédérik Boutin   
&  
Gabriel Létourneau

Travail présenté à M Yacine Benahmed

Dans le cadre du cours Projet en informatique IINF34515-TU

26 avril 2024

Table des matières

[1. Problématique 3](#_Toc164018025)

[2. Objectifs 3](#_Toc164018026)

[3. Méthodologie suivie 4](#_Toc164018027)

[3.1 - Identification des cas d’utilisation pertinents 4](#_Toc164018028)

[3.2 - Identification des sources de documents pertinents 5](#_Toc164018029)

[3.3 – Identification des données 5](#_Toc164018030)

[3.4 – Extraction des données 5](#_Toc164018031)

[3.5 – Validation des données 6](#_Toc164018032)

[3.6 - Stockage des données 6](#_Toc164018033)

[3.7 – Visualisation et mesure de l’efficacité 6](#_Toc164018034)

[4. Évaluation 6](#_Toc164018035)

[Identification des données 7](#_Toc164018036)

[Extraction des données 7](#_Toc164018037)

[Validation des données 7](#_Toc164018038)

[Stockage des données 7](#_Toc164018039)

[Visualisation et mesure de l’efficacité 7](#_Toc164018040)

[Qualité et organisation du code source 7](#_Toc164018041)

[5. Résultats 8](#_Toc164018042)

[6. Conclusion 8](#_Toc164018043)

[7. Références 9](#_Toc164018044)

[8. Annexe 10](#_Toc164018045)

# Problématique

Les documents PDF représentent un moyen courant de partage d’information par les organismes et les entreprises de toutes sortes. Il existe deux catégories d’information que peut transmettre un document PDF, soit une information qualifiée de semi-structurée avec l’utilisation de tableaux de données, ainsi qu’une information qualifiée de non-structurée représentée sous forme de cartographie et de diagrammes de tout genre. Dans le cadre de notre projet, le mandat était de développer un programme permettant la récupération automatique des informations semi-structurées, présent au sein des tableaux dans les documents PDF. Il est à noter que le mandat final de ce projet a été changé au milieu de la session, puisque le mandat initial du projet incluait le développement d’un outil automatisé de récupération de données structurées sur les différents sites web publiques des compagnies canadiennes d’électricité.

# Objectifs

Les objectifs de ce projet incluent :

* le développement d’un outil permettant l’identification automatique de l’information semi-structurée présente aux sein des tableaux dans des documents PDF;
* l’architecture de programmation de l’outil devra favoriser l’efficacité dans le traitement des données et devra favoriser le principe de séparation des préoccupations (***separation of concerns***) afin de facilité la reprise et le développement futur du projet suite à la fin de notre présent mandat;
* l’outil devra récupérer les données à l’intérieur des tableaux, leurs titres, ainsi que leurs métadonnées pertinentes;
* l’outil devra inclure un moyen de mesurer l’efficacité dans la détermination des données et dans l’intégrité de données extraites;
* l’outil devra utiliser des technologies et des moyens libres de droits d’utilisation et dit ***open source***, puisqu’il devra l’être lui-même;
* l’outil devra implémenter un mécanisme de détection des erreurs dans le cas où il y a des caractéristiques qui suggèrent un problème d’intégrité dans les données extraites d’une table, ou encore qui suggèrent une erreur dans la détermination de la table;
* l’outil devra implémenter un mécanisme de visualisation des zones de données récupérées pour facilité la détermination et la résolution des cas d’erreurs;
* l’outil devra implémenter un mécanisme de filtre afin de permettre la possibilité d’une extraction plus spécifique de l’information, en fonction de mots clés;

# Méthodologie suivie

La méthodologie comprend la segmentation du projet en différentes étapes logiques, présentées ci-bas. Il est à noter que la formule utilisée tout au long du projet a été le développement sur des durées de deux semaines avec des réunions de présentation et de discussion sur les problématiques rencontrés avec les experts du domaine et notre enseignant superviseur.

## 3.1 - Identification des cas d’utilisation pertinents

L’identification des différents cas d’utilisation pertinents c’est fait une première fois au début du projet, puis au fur et à mesure de l’expérimentation avec l’apparition des cas problématiques. Rappelons que les cas d’utilisations représentent ici des types différents de représentation des données dans les des tableaux des documents PDF. Jusqu’à maintenant, il a été possible d’identifier les cas suivants :

* tableau sous forme d’image dans le PDF;
* tableau superposé sur un filigrane;
* tableau qui possède une zone de texte à proximité de son contour;
* tableau qui possède une zone d’entêtes multiples sur plusieurs lignes;
* tableau qui se continu sur plusieurs pages;
* tableau dont l’information est difficilement lisible;
* tableau sans titre;

## 3.2 - Identification des sources de documents pertinents

Les documents PDF ayant servis au développement de ce projet ont été pris de la base de données ***CODERS-R2-Dataset Ingestion Ledger*** et nous a été distribué par les membres du ***Carrefour de Modélisation Énergétique***. Nous avons utilisé plus particulièrement documents PDF suivants pour nos tests de fonctionnement :

* 5T-14.pdf
* 7T-18.pdf
* 11-01-2017-bcuc-site-c-inquiry-final-report.pdf
* 206.3-UCAP.pdf
* data1.pdf (ne provient pas de la base de données)
* DOC\_63911\_B-1-FBC-LTERP-and-LongTerm-DSM-Plan.pdf
* fortisbc-2012-integrated-system-plan---volume-1---30june2011.pdf
* us-hydropower-market-report-full-2021.pdf

## 3.3 – Identification des données

L’identification des données se fait dans le processus **« Trouver les tables » (*Find tables*)**. Il est à noter que le processus **« Vérifier les limites d’une table » (*Find table limits*)** y joue un rôle indirect, puisqu’il permet d’identifier le lien entre les données d’une table se distribuant sur plusieurs pages. Ainsi, c’est-à-dire que ce processus permet d’obtenir le bon ordre de distribution des données. Il est possible d’obtenir les informations de ce processus au point 5 des résultats, voir également la figure 1 en annexe pour une représentation simplifiée de notre modèle sur nos processus d’affaire.

## 3.4 – Extraction des données

L’extraction des données se produit dans deux processus, soit dans le processus **« Trouver la table » (*Find table*)** qui va s’occuper de l’extraction des textes présent dans les zones de données identifiées, et dans le processus **« Trouver le titre » (*Find title*)** qui s’occupe de repérer le titre lié à une table en fonction de points de repères caractéristiques ainsi qu’en fonction des positions des tableaux avoisinants. Il est possible d’obtenir les informations de ce processus au point 5 des résultats, voir également la figure 1 en annexe pour une représentation simplifiée de notre modèle sur nos processus d’affaire.

## 3.5 – Validation des données

La validation des données des données se produit dans le processus **« Valider l’état de la table » (*Validate table state*)** qui utilise un système d’intelligence artificielle ainsi d’identifier si les données extraites du tableau présentent des indices d’un problème d’intégrité de données ou qu’il ne s’agirait pas de données d’un tableau. Il est possible d’obtenir les informations de ce processus au point 5 des résultats, voir également la figure 1 en annexe pour une représentation simplifiée de notre modèle sur nos processus d’affaire.

## 3.6 - Stockage des données

Les données résultantes sont enregistrées dans un fichier csv qui se trouve dans le répertoire ***inputs\_outputs > pdf\_outputs***. Il est possible de voir à la figure 2 de l’annexe, les résultats d’une extraction de tableau ainsi que le tableau original dans son document pdf.

## 3.7 – Visualisation et mesure de l’efficacité

Les fonctions … sont responsables de la visualisation des périmètres de données identifiées et permettent à un utilisateur de pouvoir vérifier rapidement les erreurs soupçonnées.

La mesure de l’efficacité se fait par les fonctions … . Pour le moment, l’efficacité est mesurée par rapport à des modèles qui ont été validés et certifiés par une personne. Les mesures de validations comprennent deux catégories, soit les différences de distances par rapports au périmètre d’un tableau, zone de titre ou de métadonnées, ainsi que les comparaisons en récupération de données de sortie.

Il est possible d’obtenir les informations de ce processus au point 5 des résultats, voir également la figure 1 en annexe pour une représentation simplifiée de notre modèle sur nos processus d’affaire.

# Évaluation

L’évaluation comprend un retour sur l’efficacité de fonctionnement ainsi que les contraintes et les limitations de l’outil développé. Comme pour la section précédente, l’évaluation sera segmentée en fonction des principales fonctionnalités du projet.

## Identification des données

Les tests de fonctionnements ont su démontrer que …

## Extraction des données

Les tests de fonctionnements ont su démontrer que …

## Validation des données

Les tests de fonctionnements ont su démontrer que …

## Stockage des données

Les tests de fonctionnements ont su démontrer que …

## Visualisation et mesure de l’efficacité

Les tests de fonctionnements ont su démontrer que …

## Qualité et organisation du code source

Comme il est possible de le voir à la figure 3 de l’annexe, le code est segmenté de façon logique au sein du projet. Les particularités d’une classe, propres au langage Python, nous ont orienté vers un développement en module plutôt que vers un développement en classe. Ceci permet donc une importation ciblée des différentes fonctionnalités. En effet, puisqu’il n’y a pas de restrictions d’accès aux méthodes privées des classes en langage Python, le développement en classe ne nous donne aucun avantage. Le développement en module permet d’importer uniquement les fonctions propres à nos besoins. Par exemple, un utilisateur n’ayant besoin que de la visualisation importerait uniquement la fonction …

# Résultats

Parler ici du produit finis, de la description de son architecture, de ses différentes options, fonctionnalités, autre, etc.

# Conclusion

En conclusion, …

# Références

À compléter …

# Annexe