Connexion aux fichiers PDF

Bienvenue dans cette vidéo consacrée au connecteur PDF. Vous pouvez télécharger les fichiers PDF pour suivre les étapes dans votre propre version de Tableau.

Les fichiers PDF peuvent contenir de précieuses données à analyser dans Tableau. Le connecteur PDF de Tableau est conçu pour faciliter l'importation de données contenues dans des tableaux. Étant donné que les données provenant de fichiers PDF ne comportent pas de métadonnées, il peut être nécessaire d'y apporter quelques modifications une fois la connexion effectuée.

Connexion aux tableaux de données dans les fichiers PDF

Tableau peut lire les tableaux croisés de données provenant de fichiers PDF, qui dans l'idéal se présentent comme ceci, avec des colonnes, des lignes et une seule ligne de données pour chaque ligne. Ici, les données sur le cours des actions commencent à la page 2. Connectons-nous à ces données dans Tableau. Dans le volet Connexion, sélectionnez le fichier PDF souhaité. Un outil de sélection de page s'affiche. Vous pouvez numériser le document entier, une page spécifique ou une plage de pages. Nous allons sélectionner les pages 2 à 8, puis nous allons ajouter la première page dans l'espace de travail. L'une des lignes de l'en-tête du tableau porte à confusion. Activez l'interpréteur de données pour obtenir des en-têtes corrects.

Union

Chaque page est représentée par une table unique, mais vous pouvez facilement les regrouper à l'aide d'une union, étant donné que chaque table possède la même structure et les mêmes en-têtes. Maintenant que la première table est ajoutée à l'espace de travail, faites glisser les tables des autres pages sous la table de la première page pour créer l'union. Vous obtenez une nouvelle colonne appelée « Nom de la table », et si vous faites défiler la page vers le bas, vous pouvez constater que les données issues de la page 3 s'alignent parfaitement avec celles de la page 2.

Nettoyage de tables déstructurées

Les tableaux de ce fichier PDF sont clairement structurés, ce qui n'est pas forcément le cas pour tous les fichiers PDF que vous utilisez. De manière générale, pour assurer une connexion optimale avec Tableau, le fichier doit contenir une structure tabulaire similaire à cela, avec une valeur unique par cellule de table, sans hiérarchies ni en-têtes imbriqués, ni tables imbriquées. Il n'est cependant pas nécessaire que le fichier PDF soit parfait pour être exploité. Connectons-nous à un autre fichier, qui n'est pas aussi structuré que le premier fichier. La table que nous cherchons se trouve ici, à la page 14. Attention : Tableau traite les numéros de page comme des nombres absolus, ce qui peut ne pas correspondre à la pagination du document.

Ajoutons une nouvelle source de données et choisissons la page 14. L'option « Renumériser le fichier PDF », dans le menu déroulant sous la connexion de données, permet de modifier la sélection de pages. La page ne comportait qu'un seul tableau, or trois options apparaissent dans la partie gauche. Tableau a détecté trois manières d'importer les données de ce tableau. Si vous ajoutez chaque option individuellement à l'espace de travail, vous pouvez voir ce qu'elle contient. Attention : si Tableau propose plusieurs versions de chaque tableau, lorsque vous réalisez l'union de plusieurs pages comme nous l'avons fait précédemment, veillez à utiliser la bonne version de chaque tableau, et non différentes versions du même tableau.

Faites glisser la table 1 vers l'espace de travail et utilisez l'interpréteur de données : la table 1 semble contenir toutes les informations dont nous avons besoin, mais pour une raison inconnue les années 1995 à 1997 s'affichent dans une seule colonne. Si vous souhaitez utiliser cette version des données, vous pouvez les nettoyer à l'aide d'une scission personnalisée, en scindant toutes les colonnes sur la base des espaces, puis renommer les différents champs en utilisant les années adéquates.

Voyons à quoi ressemblent les deux autres options. Il semblerait que la table 2 contienne le bas du tableau d'origine, et la table 3 la partie supérieure. J'aime la manière dont les colonnes sont réparties dans les tables 2 et 3, et je vais donc les utiliser. Premièrement, réalisez l'union de ces deux tableaux en faisant glisser la table 2 sous la table 3. La colonne F1 et « inflows » ne concordent pas. Sélectionnez-les puis choisissez l'option Fusionner des champs discordants. La colonne s'affiche maintenant comme prévu. Vous pouvez la renommer « Water Sources ».

Gestion des valeurs null

Plusieurs lignes contiennent des valeurs null, parce que le sous-en-tête, par exemple « Change in storage » a été interprété comme une ligne de données, ou parce qu'un texte a été séparé en plusieurs lignes, comme « abstraction from hydroelectricity », qui a été séparé en deux lignes différentes. Pour gérer ces valeurs null, nous allons ajouter un filtre de source de données. Cliquez sur Ajouter dans le coin supérieur droit et ajoutez un filtre. Toutes les colonnes ici contiennent des valeurs null. Je vais choisir F10, puis cliquer sur OK, Suivant, sélectionner Null et enfin Exclure. Ensuite, cliquez sur OK, puis à nouveau sur OK. Ces lignes ont maintenant disparu et seules les données restent visibles.

Cependant, je vais annuler cette action car je vois que plusieurs types de sources sont en réalité des totaux. Nous voulons créer un filtre pour les exclure en même temps que les valeurs null. Cliquez sur Filtre > Ajouter, puis sélectionnez « Water Sources ». Cochez toutes les cases correspondant à un en-tête ou à un total dans le PDF d'origine, ou à une ligne de valeur null. « Hydroelectricity » apparait deux fois, une fois avec des données et une fois avec des valeurs null. Je vais le laisser de côté pour éviter de supprimer des données. Cliquez ensuite sur Exclure, puis sur OK.

Correction des en-têtes et permutation

Il ne reste maintenant plus que les lignes avec des données, et une ligne avec des valeurs null pour « Hydroelectric », mais nous allons pouvoir la nettoyer. Nous n'avons pas d'en-têtes, sauf pour la première colonne, mais nous pouvons vérifier les valeurs dans le fichier PDF d'origine. Ces colonnes représentent simplement les années, de 1995 à 2010. Je vais accélérer cette procédure.

Nous pouvons maintenant permuter les données pour obtenir une colonne pour l'année, et une autre pour les millions de mètres cubes. Nous allons masquer la colonne « Nom de la table », choisir le type de données Date pour la colonne de l'année et le type de données Nombre (entier) pour la colonne des millions de mètres cubes. Nous avons maintenant la colonne dans son format final et nous pouvons y exlcure les valeurs null. Cliquez sur Modifier > Ajouter, sélectionnez Million Cubic Meters, puis Exclure les valeurs null.

Modification des alias des membres d'un champ

Il reste quelques problèmes à régler avec les noms des sources d'eau. Affichez le menu de ce champ, puis cliquez sur Alias. C'est ici que nous pouvons modifier l'alias des membres d'un champ donné. Cliquez deux fois sur l'alias, puis saisissez la valeur correcte.

- *« **Discharge from hydroelectricity** generation »*
- *« Groundwater »*
- *« **Abstraction for** hydroelectricity »*

Modification des groupes et des hiérarchies

Pour finir, j'ai remarqué que le tableau du PDF d'origine est davantage structuré et contient des catégories de sources d'eau. Vous pouvez recréer cette structure dans le volet Données. Cliquez sur Feuille 1. Tout d'abord, faites glisser « Million Cubic Meters » vers Mesure, car il s'agit d'une mesure.

Vous pouvez alors recréer les groupes. Ajoutez le champ « Water Sources » à l'étagère Lignes et créez un premier groupe. Maintenez la touche CTRL enfoncée et cliquez sur les membres à réunir dans le même groupe :

- *« Abstraction for hydroelectricity »*
- *« Discharge from hydroelectricity generation » et*
- *« Evapotranspiration »*
- *« To sea and net abstraction »*

Utilisez l'icône du trombone dans l'infobulle pour créer le groupe. Répétez cette procédure pour créer un autre groupe avec : *« Groundwater », « Ice », « Lakes and reservoirs », « Snow » et « Soil moisture ».*

Dans le volet Données, cliquez avec le bouton droit sur ce nouveau champ, puis sur Modifier le groupe. Cliquez d'abord sur « Precipitation » pour en faire un groupe à part, même s'il n'est composé que d'un seul élément. Vous pouvez maintenant renommer les groupes : *« Inflows », « Change in Storage » et « Outflows ».* Renommez maintenant le champ groupé « Categories ». Si vous faites glisser « Water Sources » sur le champ « Categories », vous pouvez créer une hiérarchie, que vous pourrez explorer dans la vue.

Astuces sur l'utilisation des fichiers PDF

La procédure à suivre pour nettoyer les données d'un fichier PDF après la connexion initiale peut varier et cette vidéo vous présente les outils que vous pouvez utiliser pour mieux préparer vos données à l'analyse. Rappelez-vous que Tableau rencontrera des problèmes pour traiter les fichiers PDF contenant des tables secondaires, des hiérarchies dans les en-têtes ou des lignes contenant plusieurs lignes devant être interprétées comme des lignes uniques. Pour finir, notez que les couleurs et la trame de fond peuvent avoir une influence sur la manière dont les données sont interprétées, car Tableau analyse les fichiers PDF à la recherche de cellules et de tableaux de données.

Conclusion

Merci d'avoir suivi cette vidéo de formation sur le connecteur PDF. Nous vous invitons à découvrir les autres vidéos de formation gratuite pour en apprendre davantage sur l'utilisation des produits Tableau.