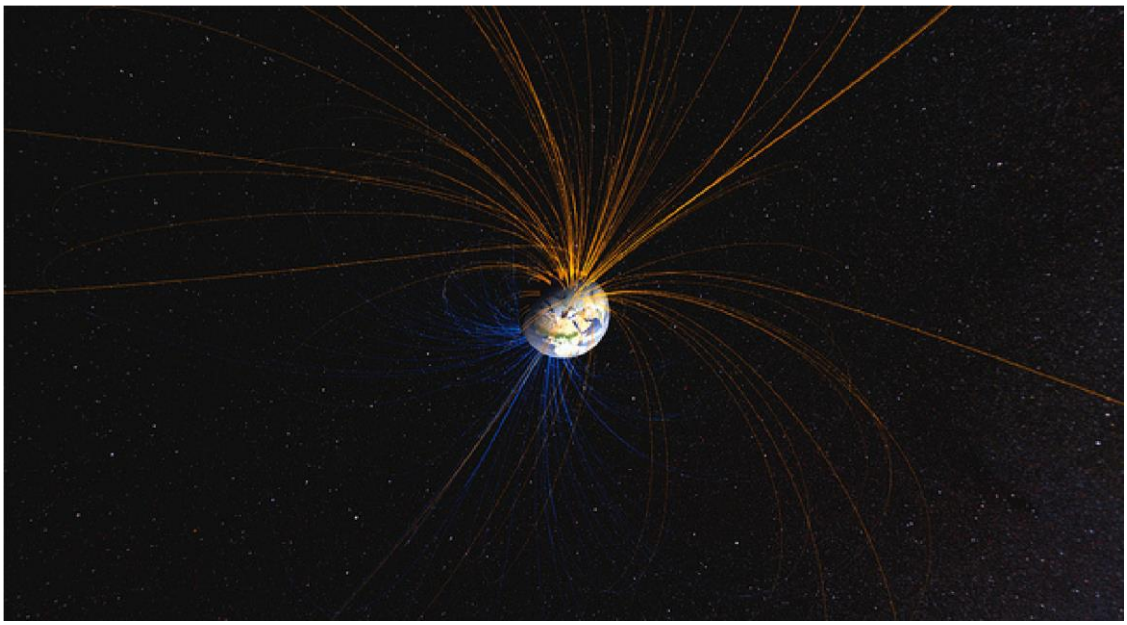
Gestion des Données Ouvertes en Agriculture et Nutrition

Ce cours en ligne est le fruit d'une collaboration entre les partenaires de GODAN Action, y compris Wageningen Environmental Research (WUR), AgroKnow, AidData, l'Organisation des Nations Unies pour l'Alimentation et l'Agriculture (FAO), le Forum Mondial sur la Recherche Agricole (GFAR), l'Institut des Etudes du Développement (IDS), le Land Portal, l'Open Data Institute (ODI) et le Centre Technique de Coopération Agricole et Rurale (CTA).

GODAN Action est un projet de trois ans du Département pour le Développement International du Royaume-Uni pour permettre aux utilisateurs, producteurs et intermédiaires de données de s'engager efficacement avec les données ouvertes et maximiser leur potentiel d'impact dans les secteurs de l'agriculture et de l'alimentation. Nous travaillons en particulier à renforcer les capacités, à promouvoir des normes communes et les meilleures pratiques et à améliorer la manière dont nous mesurons l'impact. [www.godan.info]

MODULE 3: RENDRE LES DONNÉES OUVERTES

LEÇON 3 .2: Gestion dynamique des jeux de données



Objectifs et résultats d'apprentissage

Cette leçon a pour objectif ;

- De donner un aperçu des procédés automatisés de publication des données et de saisie/agrégation manuelle et donner des exemples
- D'introduire des moyens simples d'assurer la cohérence des données.

À la fin de cette leçon, vous devrez être en mesure de :

- Comprendre les procédés qui jouent un rôle dans le maintien et l'exposition des ensembles de données dynamiques
- Connaître des exemples existants d'ensembles de données dynamiques
- Connaître les méthodes de vérification de la cohérence des données dans les ensembles de données dynamiques.

Sommaire

Liste des illustrations

1. Introduction

Dans la leçon précédente, nous avons examiné les principes de gestion des données et son large éventail d'activités, des aspects administratifs aux aspects techniques. Les données peuvent être publiées de diverses façons, selon que les données sont recueillies et regroupées manuellement ou au moyen de procédés automatisés, ce qui a une incidence sur leur publication.

Il est important de comprendre les différences entre les deux types d'ensembles de données dont il sera question dans le présent module - statiques et dynamiques - ainsi que les procédés de gestion qui s'y rapportent. Les progrès technologiques que nous constatons aujourd'hui ont mené à la mise au point d'outils que nous pouvons utiliser pour donner accès aux données, et nous donnerons des exemples d'outils qui peuvent être utilisés pour exposer des données dynamiques. Ce module décrira également des routines simples de cohérence des données à utiliser pour contrôler et vérifier les jeux de données dynamiques.

2. Jeux de données dynamiques et statiques

Les données dynamiques désignent les données qui sont modifiées au fur et à mesure que d'autres mises à jour deviennent disponibles[1]. Au contraire, les données statiques, également appelées données persistantes, sont moins consultées et ne sont pas susceptibles d'être modifiées. Les données dynamiques sont différentes des données en continu (streaming) en ce sens qu'il n'y a pas de flux constant d'information ; les mises à jour peuvent survenir à tout moment, avec des périodes d'inactivité entre les deux.

En informatique, les structures de données statiques sont dotées d'une zone de mémoire fixe dans laquelle elles peuvent fonctionner. Il n'est pas possible d'étendre cette taille fixe en cours d'exécution. Par conséquent, les emplacements de chaque élément sont fixes et connus par le programme. Les structures de données dynamiques, en revanche, disposent d'une zone flexible où elles peuvent fonctionner. Il est possible d'agrandir ou de réduire la zone selon les besoins, en ajoutant ou en supprimant des éléments de la structure de données. Il serait donc inefficace d'utiliser des structures dynamiques pour stocker des ensembles de données qui ne changeront pas. Dans ce cas, l'utilisation de structures de données statiques permet d'économiser les ressources du système et d'accéder plus rapidement aux éléments. Les utilisateurs ou les développeurs sont responsables de l'utilisation de structures de données appropriées, selon la situation.

[1] https://en.wikipedia.org/wiki/Dynamic_data

L'essor des technologies de précision en agriculture modifie la façon dont les agriculteurs gèrent leurs terres et leur bétail, comme les systèmes de géo-positionnement par satellite et les capteurs qui détectent les nutriments et l'eau dans le sol. Ces technologies aboutissent finalement à la collecte de données plus dynamiques, qui sont traitées automatiquement.

Ensuite, nous examinerons les différents méthodes et procédés utilisés pour la publication manuelle ou automatique des données.

2.1. Publication des données - saisie manuelle

2.1.1. Publication sous forme de fichiers sur un site Web

Les ensembles de données statiques peuvent être partagés via des canaux statiques tels que des sites Web. Certaines organisations n'ont que quelques ensembles de données à partager, ce qu'elles font en publiant les fichiers sur leur site Web. La solution logicielle utilisée pourrait être n'importe quel système de gestion de contenu (CMS), tel que Drupal, Wordpress, en combinaison avec une base de données interne. Par exemple, la ville néerlandaise de Hengelo a inclus une page Web pour partager les données sur le site Web de la ville, comme le montre l'illustration 1.

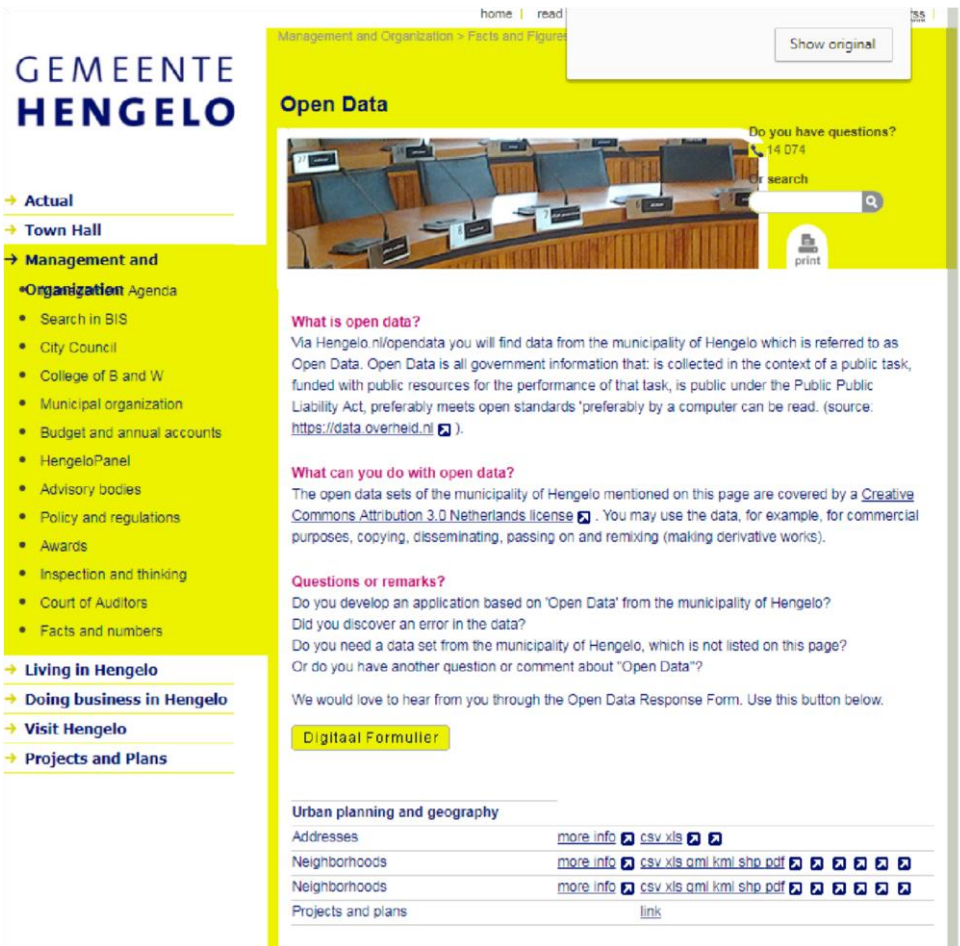


Illustration 1 Le site Web Hengelo Open Data http://www.hengelo.nl/opendata

2.1.2. Téléchargements de portails Web

Le téléchargement sur un portail est l'un des canaux les plus utilisés pour publier des données ouvertes. L 'illustration 2 ci-dessous montre une capture d'écran du portail CKAN sur lequel les données sont téléchargées manuellement[2]. Cet interface CKAN comporte certaines restrictions, telles que l'absence de réutilisation de modèles similaires et de support multilingue.

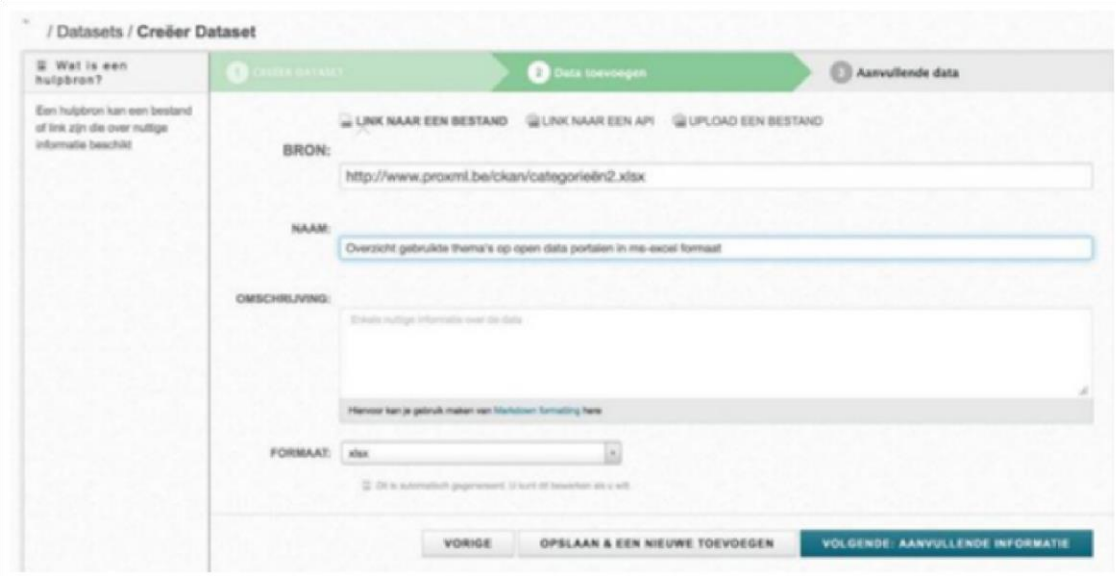


Illustration 2 Exemple d'un ensemble de données ajouté manuellement au portail de données flamand à l'aide de l`interface CKAN[3]

Un portail fournit, via un catalogue de métadonnées, un point d'accès unique aux données.

Voici des exemples de portails de données ouvertes contenant des données sur l'agriculture :

- EUROPE : https://data.europa.eu/euodp/en/data
- ETATS-UNIS : https://www.data.gov/
- ROYAUME UNI : https://data.gov.uk/
- FAO: http://www.fao.org/data/en/
- Banque mondiale : http://data.worldbank.org/
- Mettre fin à la faim en milieu rural (ERH) : https://endingruralhunger.org/
- CGIAR: http://www.cgiar.org/resources/open/data-managementsystem/
- Food Aid – WFP : le Système d'information sur l'aide alimentaire (FAIS) met à disposition des données détaillées sur l'aide alimentaire mondiale. FAIS est accessible à l'adresse suivante : http://www.wfp.org/fais/.
- Nutrition animale - FAO : les données sur la composition nutritionnelle des aliments pour animaux sont disponibles à l'adresse suivante : http://www.feedipedia.org/

[2] https://www.europeandataportal.eu/en/providing-data/goldbook/publishing-data
[3] https://www.europeandataportal.eu/en/providing-data/goldbook/publishing-data

- Pêche - FAO : les liens vers les publications contenant des données sur la pêche se trouvent à l'adresse suivante : http://www.fao.org/fishery/statistics/en
- Alimentation et nutrition : pour des liens sur l'alimentation et la nutrition dans ce site Web, cliquez ici.

Tous les portails de données ouvertes (2500+) sont listés ici : https://www.opendatasoft.com/a-comprehensive-list-of-all-open-dataportals-around-the-world/

California Open Data - exemple d'un flux manuel pour la publication de données sur un portail de données ouvertes

La publication de données ouvertes est contrôlée par un flux séquentiel. Un coordonnateur des données se voit attribuer l'un de ces deux rôles :

- Créateur de contenu, contributeur de workflow : ce rôle est attribué aux spécialistes des données au niveau départemental qui créent et téléchargent les ensembles de données.
- Éditeur, modérateur de workflow : ce rôle est assigné à un gestionnaire de données au niveau du service qui examine et approuve les données avant leur diffusion.

Le processus de publication est décrit en détail à l'adresse suivante : https://office-of-digitalinnovation.github.io/data-training/publishing/.

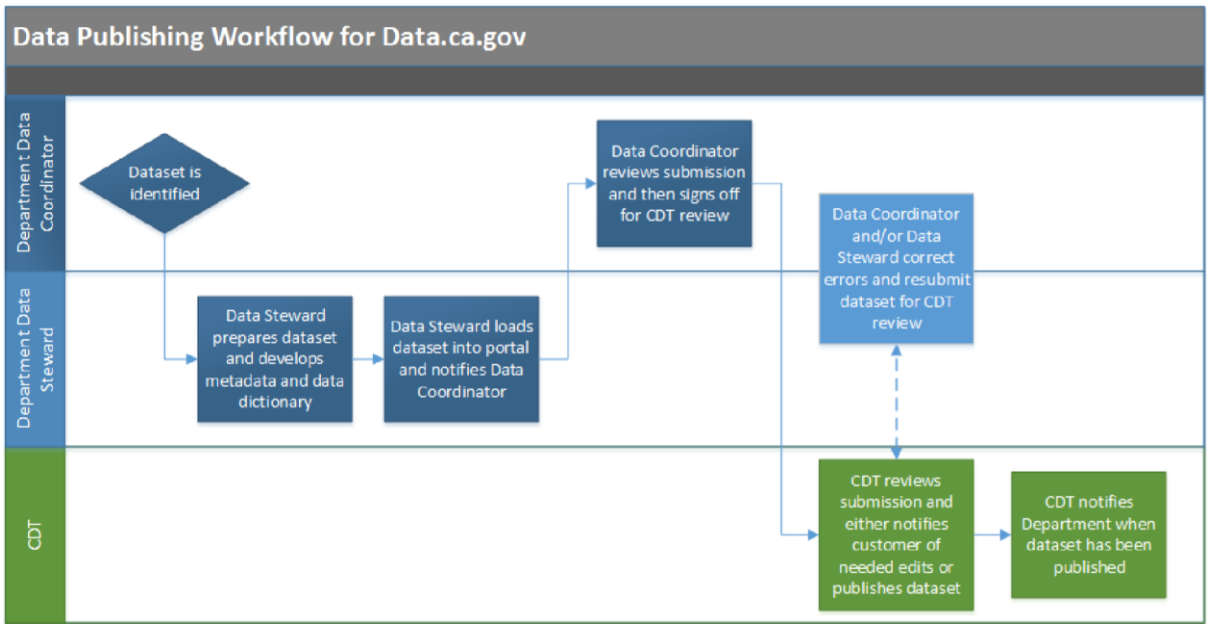


Illustration 3 California Open Data fournit un exemple de flux manuel pour la publication sur un portail de données ouvertes.

2.2. Publication de données - processus automatisés

Sources des données automatisées. Au fur et à mesure que l'Internet des objets (IdO) se développe, en particulier avec le développement de capteurs bon marché, l'automatisation des possibilités de collecte de données s'accroît. Grâce à des capteurs précis, les agriculteurs peuvent recueillir automatiquement des données sur le temps, le sol, la qualité de l'air et la maturité des cultures, ce qui leur permet de prendre des décisions plus judicieuses. Des appareils tels que les téléphones mobiles sont également utilisés pour collecter des données passives : par exemple, l'application mobile WeatherSignal[4] utilise des capteurs téléphoniques de la région, qui sont ensuite affichés sur leur carte météo mise à jour en temps réel, afin de mesurer les conditions atmosphériques. Les capteurs des villes, des maisons, des voitures, des tours de téléphonie cellulaire et même des téléphones mobiles peuvent maintenant fournir des données d'observation qui pourraient également être utilisées dans les modèles de prévision. À mesure que les ensembles de données automatisés continuent de croître de façon exponentielle, de nouvelles technologies ont été introduites pour une diffusion plus durable.

L'automatisation des données est le procédé de mise à jour des données sur votre portail de données ouvertes par programme, plutôt que manuellement. L'automatisation du procédé de téléchargement des données est importante pour la durabilité à long terme de votre programme de données ouvertes. Toute donnée mise à jour manuellement risque d'être retardée car il s'agit d'une tâche supplémentaire qu'une personne doit accomplir dans le cadre du reste de sa charge de travail.

L'automatisation de la publication des données fournira à vos utilisateurs des mises à jour régulières et prévisibles et permettra d'améliorer l'efficacité du processus de publication. Il y a trois éléments communs à l'automatisation des données - l'Extraction, la Transformation et le Chargement.

- *L'Extraction* : le procédé d'extraction de vos données à partir d'un ou plusieurs systèmes sources
- *La Transformation* : le procédé de transformation de vos données dans la structure nécessaire, comme un format de fichier plat comme un CSV ; cela pourrait aussi inclure des opérations comme changer toutes les abréviations de États pour en indiquer le nom complet
- *Le Chargement* : le processus de chargement des données dans le système final, en l'occurrence le portail de données ouvertes.

Les procédés de publication automatisés étaient auparavant très personnalisés, les éditeurs écrivant un script ETL Extraction–Transformion–

[4] https://play.google.com/store/apps/details?id=com.opensignal.weathersignal&hl=en

L'automatisation des processus de chargement des données demandait jusqu'à présent beaucoup de travail, les auteurs devant écrire un script de chargement (ETL) à partir de zéro. Les procédés ETL effectuent ces trois tâches : extraire les données brutes de la source originale, les transformer en un format plus utile et les charger dans le portail de données ouvertes. Chacun de ces processus est essentiel à l'automatisation complète de vos téléchargements de données avec succès. L'illustration 4 décrit un procédé ETL standard.

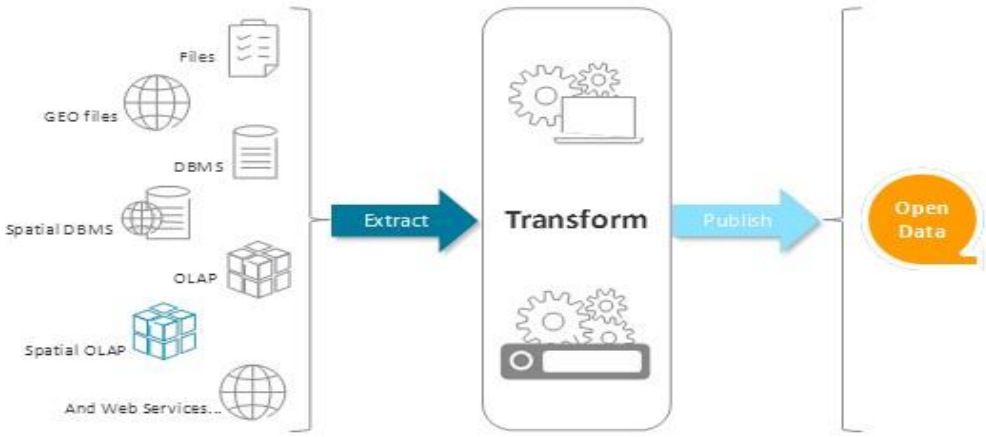


Illustration 4 Un procédé ETL : il extrait les données brutes de la source originale, les transforme en un format plus utile et les charge/publie dans le portail de données ouvertes[6]

Stratégie d'automatisation des données. Tout d'abord, il est important de déterminer une stratégie générale d'automatisation des données pour votre organisation. Avoir une stratégie à l'avance vous aidera à engager les bonnes personnes, les bons outils, au bon moment au sein de votre organisation.

Identifier qui est responsable de l'automatisation des données dans votre organisation. Différents groupes seront responsables de différentes parties du procédé ETL :

Centralisé : le service informatique central est responsable de l'ensemble du procédé ETL et de l'automatisation des données :

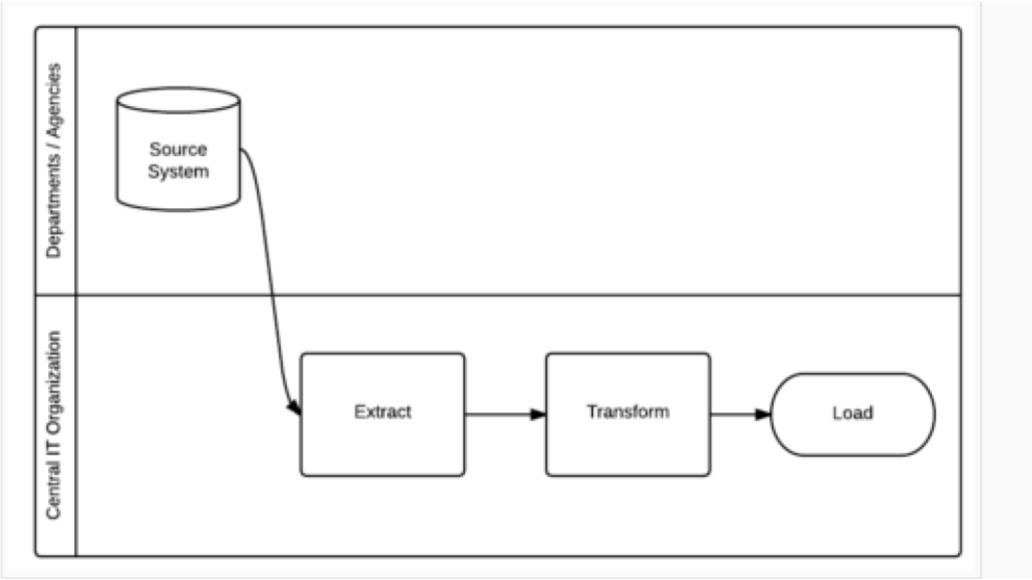


Illustration 5 Un procédé ETL centralisé

Hybride : ce modèle peut varier, mais il arrive souvent que les différents organismes ou ministères soient responsables des procédés d'extraction et de transformation, et que le service central de TI soit responsable du procédé de chargement (illustration 6) :

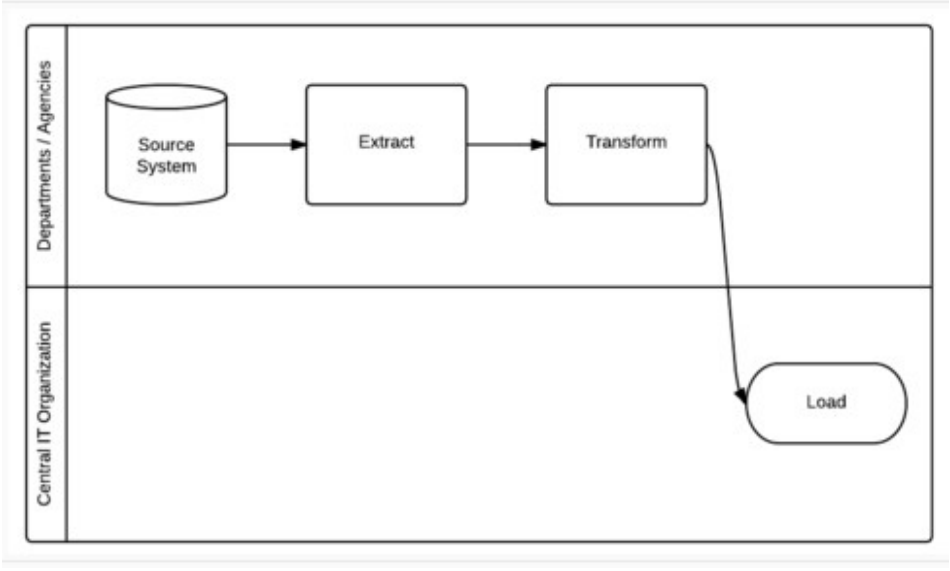


Illustration 6 Un procédé ETL hybride

Décentralisé : les agences/services individuels auront chacun leur propre procédé ETL (Illustration 7).

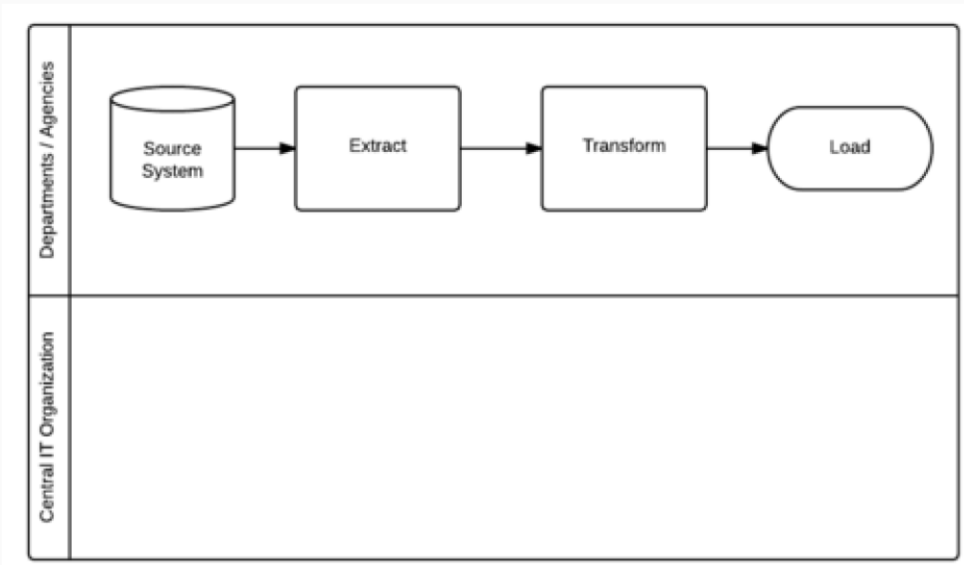


Illustration 7 Un procédé ETL décentralisé

Une fois que vous aurez défini où les étapes ETL surviennent au sein de l'organisation, vous devrez déterminer où l'automatisation s'inscrira dans votre processus de publication. Lorsque vous créez votre shéma de publication, assurez-vous de clarifier les étapes d'extraction, de transformation et de chargement, ainsi que les personnes qui seront responsables de ces étapes.

Sélection des données à automatiser. Plus vous adopterez une approche " d'automatisation par défaut " pour télécharger des données, moins vous aurez besoin de ressources à long terme pour maintenir une qualité de données élevée. Voici quelques conseils pour trouver des ensembles de données pertinentes pour un téléchargement automatique :

- L'ensemble de données est-il mis à jour trimestriellement ou plus fréquemment ?
- Y a-t-il des transformations ou des manipulations qui doivent être effectuées sur l'ensemble de données avant le téléchargement ?
- L'ensemble de données est-il volumineux (plus de 250 Mo) ?
- Pouvez-vous seulement obtenir les lignes modifiées pour chaque mise à jour ultérieure (plutôt que le fichier complet) ?
- Est-il possible d'obtenir des données du système source plutôt que d'un individu ?

Les ensembles de données qui reçoivent un " oui " fort à l'une ou l'autre des questions ci-dessus sont d'excellents candidats pour l'automatisation des mises à jour, car l'automatisation peut éliminer le risque d'un manque de temps et de ressources pour préparer manuellement les jeux de données.

Étapes de l'automatisation : Une fois que vous aurez visualisé l'ensemble des éléments d'automatisation des données au sein de votre organisation, vous pouvez commencer à mettre en œuvre votre stratégie. Voici quelques étapes pour commencer :

1. *Identification des données* : choisissez-en un ou deux jeux de données de grande valeur où l'accès aux systèmes sources sera facile (c'est-à-dire commencez par le " fruit à portée de main ").
2. Reportez-vous à votre inventaire pour déterminer les systèmes sources auxquels vous avez déjà accès.
3. *Détermination de l'accès* : Déterminez comment le ministère ou l'organisme concerné obtiendra les données. S'agira-t-il d'une requête SQL, du téléchargement d'un CSV, etc. ? Le responsable des données serait probablement la meilleure ressource pour accéder au système source d'un ensemble de données.
4. *Définition des transformations :* décrivez toutes les transformations souhaitées pour cet ensemble de données. Cela peut être aussi simple que de changer des acronymes complexes en noms en texte intégral, ou aussi compliqué que de transformer une base de données relationnelle en un fichier plat-CSV.
5. Travaillez avec le responsable des données et le responsable de la conservation des données pour *comprendre quels champs* doivent être tirés et comment ils doivent être formatés pour la publication.
6. *Développez et testez le procédé ETL :* sur la base des exigences définies aux étapes 2 et 3, sélectionnez un outil de publication ETL et publiez l'ensemble de données sur le portail de données ouvertes. Confirmez que le jeu de données a été chargé ou mis à jour avec succès tout au long du processus.
7. *Calendrier* : Planifiez votre ensemble de données pour des mises à jour en temps opportun.
8. Référez-vous aux *champs de métadonnées* sur la collecte de données, la fréquence de réactualisation, et la fréquence de mise à jour.

3. Cohérence des données

Des données cohérentes sont des données qui sont techniquement correctes et adaptées à l'analyse statistique. Il s'agit de données qui ont été vérifiées pour les valeurs manquantes, les valeurs spéciales, les erreurs (évidentes) et les valeurs extrêmes, qui sont soit supprimées, corrigées ou imputées. Les données

sont conformes aux contraintes fondées sur les connaissances du monde réel sur le sujet que les données décrivent[5].

La qualité des données est évaluée en appliquant des procédures de vérification et de validation dans le cadre du processus de contrôle de la qualité. La vérification et la validation sont des éléments importants de la gestion des données qui aident à assurer la validité et la fiabilité des données. L'Environmental Protection Agency des États-Unis définit la vérification des données comme le processus d'évaluation de l'exhaustivité, de l'exactitude et de la conformité d'un ensemble de données aux procédures requises pour s'assurer que les données sont bien celles qu'elles sont censées être. La validation des données fait suite à la vérification des données et comprend l'évaluation des données vérifiées pour déterminer si les objectifs de qualité des données ont été atteints et les raisons de tout écart.

Les principes de qualité des données doivent être appliqués à toutes les étapes du processus de gestion des données (saisie, numérisation, stockage, analyse, présentation et utilisation). Il y a deux clés pour améliorer la qualité des données : la prévention et la correction. La prévention des erreurs est étroitement liée à la fois à la collecte des données et à leur saisie dans une base de données. Bien que des efforts considérables puissent et doivent être consacrés à la prévention des erreurs, il n'en demeure pas moins que des erreurs continueront d'exister dans de grands ensembles de données et que la validation et la correction des données ne peuvent être ignorées.

3.1.1. Listes de contrôle pour évaluer la cohérence de vos données

1. Votre ensemble de données est-il complet ? Chaque ensemble de données devrait :

- Contenir une ligne d'en-tête avec une seule description de ce qui est affiché. Cela signifie qu'une fois qu'une structure d'ensemble de données est en place, elle ne devrait pas changer lorsque les sources sont ajoutées. Dans les métadonnées, l'en-tête doit être décrit.
- Etre étiqueté avec un numéro de version. Une fois la mise à jour effectuée, l'ensemble de données devrait recevoir un nouveau numéro de version afin que l'auditoire puisse suivre les changements.
- Contenir des informations sur son origine. Sur quoi portent les données, d'où proviennent-elles et dans quel but ont-elles été publiées ?
- Recevoir un statut : Brouillon, validation, définitive

2. Les données sont-elles curées ? Vérifiez les aspects suivants:

- champs vide
- Données factices et valeurs par défaut - sont-elles correctes ?
- valeurs incorrectes

[5] E de Jonge and M van der Loo, 2013, *An introduction to data cleaning with R* Statistics Netherlands, The Hague, Netherlands. Available at: https://cran.r project.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf

- entrées doubles
- renseignements confidentiels

Différents exemples d'erreurs et d'incohérences dans les données, et comment les corriger avec la validation des schémas et des outils plus avancés comme Open Refine[6] sont traités en détail dans le module 2, leçon 2.2.

3. Les données sont-elles exactes ? Vos données sont-elles exactes ? Les aspects les plus importants concernant la précision sont les suivants :

- Les données sont-elles suffisamment précises pour l'usage auquel elles sont destinées ?
- Leur exactitude affecte-t-elle leur fiabilité ?
- Les choix concernant l'intervalle sont-ils décrits ?
- Les données doivent-elles être agrégées ou désagrégées ?

La documentation est la clé d'une bonne qualité des données. Sans une bonne documentation, il est difficile pour les utilisateurs de déterminer l'aptitude à l'utilisation des données et pour les responsables de savoir ce qui a été fait, et par qui les contrôles de qualité des données ont été effectués. La documentation est généralement de deux types, et il faut en tenir compte dans la conception de la base de données. La première est liée à chaque enregistrement et enregistre les vérifications de données qui ont été effectuées, les changements qui ont été apportés et par qui. La seconde est la métadonnée qui enregistre l'information au niveau de l'ensemble de données.

4. Exposition d'ensembles de données dynamiques

4.1 API et données ouvertes

Une API (Interface de programmation d'application) est un nom commun qui permet de relier des systèmes via une interface de programmation. Une API peut être utilisée pour rendre les données ouvertes disponibles en fournissant à un ré-utilisateur un accès direct aux données ouvertes du fournisseur.

Les données contiennent généralement un ensemble spécifique de champs ou de colonnes et de codes, que les ré utilisateurs doivent comprendre pour travailler efficacement avec les données. S'il est servi par le biais d'une API ou d'un service, les ré utilisateurs devront comprendre le fonctionnement de l'API[7] pour les utiliser.

L'Open Data Institute suggère que la documentation technique que vous devez fournir avec une API comprenne les éléments suivants :

[6] http://openrefine.org
[7] https://theodi.org/guides/engaging-reusers

- **La documentation sur les formats** de données que vous fournissez, y compris éventuellement des schémas pour tous les vocabulaires que vous utilisez ;
- **Des listes de codes** qui fournissent plus de détails sur chacun des codes utilisés dans vos données ; une façon de fournir cette information est d'avoir une URL qui fournit de la documentation sur chaque code et d'établir un lien vers cette URL dans les données ;
- **Une documentation de service** qui décrit le fonctionnement de toute API que vous fournissez ; cela peut inclure des liens vers des descriptions de service lisibles par machine, le cas échéant.

Munis de cette information, les ré-utilisateurs devraient être en mesure de comprendre les données que vous publiez et comment créer des applications qui les utilisent.

4.1.1. Exemple d'API en agriculture

Portail de la sécurité alimentaire[8] : ce site de l'IFPRI[9] contient plus de 40 indicateurs liés à la sécurité alimentaire, aux prix des produits de base, à l'économie et au bien-être humain. La plupart de ces données sont disponibles pour tous les pays du monde et remontent à plus de 50 ans. Ils s'appuient sur des sources de données publiques faisant autorité comme la Banque mondiale, la FAO, l'UNICEF et d'autres, ainsi que sur leurs propres données.

Afin de rendre les données contenues sur le site aussi utiles que possible, elles peuvent être téléchargées gratuitement par le biais de l'API de données et être agrégées, mises en cache et partagées. Le portail est conçu pour regrouper ces informations de manière structurée et vérifier la qualité et la pertinence des données.

4.1.2. Exemple d'API de données météorologiques

OpenWeatherMap[10] : Le site publie des données météorologiques ouvertes par le biais d'une API destinée aux développeurs, ce qui facilite l'intégration de l'information météorologique dans diverses applications, y compris les applications météorologiques Web et mobiles, et des solutions pour l'assurance, l'agriculture, le sport et de nombreux autres domaines.

Le service OpenWeatherMap collecte les données des stations météorologiques professionnelles et privées. Aujourd'hui ils ont plus de 40.000 stations météo ; la plupart sont des stations professionnelles qui sont installées dans les aéroports, les grandes villes, etc.

[8] http://www.foodsecurityportal.org
[9] http://www.ifpri.org
[10] http://www.openweathermap.org

4.1.3. La technologie derrière les API

Les API sont pilotées par un ensemble de technologies spécifiques, ce qui les rend facilement compréhensibles par les développeurs. Ce type de focalisation signifie que les API peuvent fonctionner avec n'importe quel langage de programmation commun, l'approche la plus populaire pour fournir des API Web étant REST (REpresentational State Transfer).

REST tire parti des mêmes mécanismes Internet que ceux qui sont utilisés pour visualiser les pages Web régulières, ce qui lui donne de nombreux avantages qui peuvent aboutir à des développements plus rapides et plus faciles à comprendre et à utiliser par les développeurs. Les API REST vous permettent de prendre les données et les fonctionnalités qui peuvent déjà être disponibles sur votre site Web et de les rendre disponibles via une API programmatique que les applications Web et mobiles peuvent utiliser. Ensuite, au lieu de représenter l'information en HTML comme le ferait un site Web, une API renvoie des données dans l'un des deux formats suivants :

- Extensible Markup Language (XML), ou
- JavaScript Object Notation (JSON).

Les développeurs peuvent alors récupérer ces données et les utiliser dans des applications web et mobiles. Cependant, XML et JSON sont facilement utilisables avec les tableurs et autres outils que les non-développeurs peuvent choisir, ce qui rend les APIs accessibles à tous.

Résumé

Les données peuvent être publiées de diverses façons, selon que les données sont recueillies et agrégées manuellement ou au moyen de processus automatisés.

Les données dynamiques désignent les données qui sont modifiées de manière synchrone au fur et à mesure que d'autres mises à jour sont disponibles.

Les jeux de données peuvent être publiés par le biais de sites Web, de portails ou d'API.

L'automatisation des données est le processus de mise à jour des données sur votre portail de données ouvertes par programmation, plutôt que manuellement. L'automatisation du processus de téléchargement des données est importante pour la pérennité à long terme de votre programme de données ouvertes.

Les trois étapes communes de l'automatisation des données sont l'extraction, la transformation et le chargement, ou ETL :

- L'extraction : le processus d'extraction de vos données à partir d'un ou plusieurs systèmes sources.
- La transformation : le processus de transformation de vos données dans la structure nécessaire, tel qu'un format de fichier plat comme un CSV.
- Le chargement : le processus de chargement des données dans le système final.

Les données cohérentes sont des données qui sont techniquement correctes et qui sont aptes à l'analyse statistique. Les données ont été vérifiées pour détecter les valeurs manquantes, les valeurs spéciales, les erreurs (évidentes) et les valeurs aberrantes sont supprimées, corrigées ou imputées. Les principes de qualité des données doivent être appliqués à toutes les étapes du processus de gestion des données (saisie, numérisation, stockage, analyse, présentation et utilisation). Lorsque vous examinez la cohérence de vos données, vous devez vérifier que votre ensemble de données est complet, les données sont propres et les données sont exactes.

Une API (interface de programmation d'application) est un nom collectif pour relier des systèmes via une interface de programmation. Une API peut être utilisée pour rendre les données ouvertes disponibles en fournissant à un ré utilisateur un accès direct aux données ouvertes du fournisseur.

Dans la prochaine leçon 3.3 *Création et gestion de dépôts de données ouvertes*, nous explorerons les différentes options disponibles pour héberger un dépôt de données ouvert et comment le gérer.

Lecture complémentaire

- Evans, S. (undated) Robots Set to Transform the Automotive and Agricultural Industries – Interview with Dr Robert Fitch (Australian Centre for Field Robotics). Available at: http://marketclarity.com.au/acfr-robots-setto-transform-the-automotive-and-agricultural-industries/
- GODAN Global Data Ecosystem Publication http://www.godan.info/documents/data-ecosystem-agriculture-and-food
- Mill, E. (2016) An Introduction to Open Data and APIs (video). DigitalGov. Available at: https://www.youtube.com/watch?v=taTdJ6oOZX4