Automatisation d'un processus de contrôle qualité de données au format tableur issues de Prodinra

Sylvain Cariou, Alexandra Coppolino, Lise Frappier

Automatisation d'un processus de contrôle qualité de données au format tableur issues de Prodinra

Sylvain Cariou[1], Alexandra Coppolino[2], Lise Frappier[1]

Résumé. Les professionnels de l'Information Scientifique et Technique (IST), de l'Inra veillent à ce que les données disponibles dans l'archive ouverte ProdInra soient de qualité. Ce travail est indispensable pour l'obtention de listes de publications et d'indicateurs les plus fiables possible. Dans ce cadre, nous avons travaillé en collaboration avec un informaticien sur une automatisation d'un processus de contrôle qualité de données extraites au format tableur pour gagner en efficacité. Cet article décrit notre démarche. Il vous propose de faire de même pour votre corpus bibliographique issu de ProdInra. Vous trouverez également des conseils sur les précautions à prendre pour utiliser les fichiers de sortie. Enfin, nous proposons des évolutions en prévision du passage de ProdInra à HaL[3].

Mots clés : valorisation, qualité des données, information scientifique et technique, publication scientifique, bibliométrie

Abstract. The INRA professionals of Technical and Scientific Information (IST) look after the quality of data available in the open repository ProdInra. This work allows obtaining some lists and indicators the most reliable possible about the publications. In this scope, we work together with a computer expert on the automatization to set up an automatized quality control process based on a extracted excel file, which would be more efficient. This paper describes the process in detail, so this process could be reused for anyone who works with a corpus of publications in ProdInra. You will find as well some precautionary advices and cares about the files that are used in the process. We propose then possible evolutions for the future ProdInra when migrated in the HAL infrastructure.

Keywords: reuse, data quality, technical and scientific information, scientific publication, bibliometrical indicators

Introduction

L'archive ouverte institutionnelle des publications des chercheurs de l'Inra est ProdInra[4] . Cette source est utilisée pour produire des indicateurs sur la production scientifique et des listes bibliographiques pour les évaluations individuelles et collectives. Pour que les indicateurs soient de bonne qualité, les données doivent être uniformisées grâce à un travail de vérification et de correction[5].

Dans la situation décrite ici, un chantier qualité sur les données de ProdInra concerne un corpus de plusieurs centaines notices, exportées dans un tableur Excel (151 colonnes = ensemble des champs de la notice ; une ligne = une notice). Les professionnels de l'IST vont vérifier un ou plusieurs champs selon le chantier. Par exemple, ils contrôlent que le champ « Comité de lecture » soit correctement renseigné selon la revue pour chaque article publié.

[1] *SMART-LERECO, Agrocampus ouest, Inra, 35000, Rennes, France sylvain.cariou@inra.fr lise.frappier@inra.fr (auteur de correspondance)*
[2] *CEE-M, Inra, CNRS, Montpellier Supagro, Université de Montpellier, 34000, Montpellier, France alexandra.coppolino@inra.fr*
[3] *Hal : Archive ouverte nationale maintenue par le CCSD, retenue comme future archive ouverte de l'INRA https://hal.archives-ouvertes.fr/*
[4] *ProdInra : Archive ouverte institutionnelle de l'INRA - http://prodinra.inra.fr*
[5] *Plus d'information sur la qualité dans ProdInra : Enjeux et mise en œuvre de chantiers qualité des données dans l'archive ouverte de l'Inra : ProdInra. Tang-Chaupitre C., Fouché S., Batifol-Garandel V., Gautret M., Le Hénaff D. Cahier des Techniques de l'INRA, Numéro spécial 2012 pp. 133-140 http://prodinra.inra.fr/record/178325*

Sylvain Cariou, Alexandra Coppolino, Lise Frappier

Un chantier qualité demande la mobilisation d'un temps variable pour les professionnels de l'IST, généralement long d'un ou plusieurs mois, selon le corpus de notices à corriger, les champs plus ou moins difficiles à analyser, la disponibilité des professionnels de l'IST impartie à cette mission.

Certaines de ces tâches sont répétitives et communes d'un corpus à un autre. C'est partant de ce constat qu'une collaboration entre métiers a débuté. Ainsi, deux étapes subsistent : l'export et la correction des données dans ProdInra.

Sans outil de contrôle automatisé les professionnels de l'IST doivent effectuer toutes les six tâches ci-dessous manuellement. Avec l'outil, quatre d'entre elles sont tout ou partie automatisables.

Exemple de chantier qualité manuels dans ProdInra – SAE2

1 Exporter le corpus depuis ProdInra à partir d'une requête spécifique, au format .XLS

2 Créer des tableaux croisés dynamiques
Ex. de chantier : sur les champs « Titre de la revue », « ISSN » de la revue et « Comité de lecture » (valeurs Oui/Non)

3 Pour notre ex., vérifier sur le site de la revue ou/et sur les listes HCERES si la validation par les pairs ou peer-reviewing est pratiquée

4 Repérer les incohérences,
Ex : des articles publiés dans une même revue, dont le champ « Comité de lecture » est renseigné différemment selon les notices

5 Générer un fichier .XLS de traitement des notices erronées avec les valeurs corrigées

6 Corriger les notices erronées dans ProdInra

Étapes automatisables grâce au Job Talend

Le fruit d'une collaboration entre métiers

La volonté de développer cet outil fait suite au développement d'un autre outil qui permet d'automatiser la production d'une note contenant des indicateurs de production synthétiques pour l'Unité SMART-LERECO[6] (économie) qui ne sera pas décrit ici.

La collaboration s'est faite entre un informaticien / gestionnaire de bases de données SAE2[7] et deux documentalistes du Département SAE2. La base de la discussion est simple : « un traitement de fichier que l'on fait deux fois de façon identique, c'est une fois de trop ».

Des compétences en codage alliées à une bonne connaissance de la base de données ProdInra ont permis de créer un outil efficace, simple à prendre en main. Celui-ci répond à des besoins essentiels de qualité des données, dans le contexte de nos missions d'appui aux chercheurs ou la recherche.

Il effectue les tâches répétitives et communes d'un corpus à un autre à la place des documentalistes. Ceux-ci n'ont plus qu'à réaliser l'analyse intellectuelle des résultats fournis par l'outil. Si l'on reprend l'exemple dans l'introduction de chantier qualité sur les revues à comité de lecture, l'outil peut effectuer les tâches 2, 3 et 5.

Le développement de l'outil d'automatisation du contrôle qualité

Le logiciel utilisé pour mettre en œuvre l'automatisation du contrôle de la qualité des données de ProdInra est Talend Open Studio[8] for Data Integration. Ce logiciel gratuit fait partie de la famille des ETL (Extract Transform and Load) qui permet de réaliser des extractions depuis une source de données, de transformer ces données pour les importer vers une autre cible de données. La particularité de Talend Open Studio est de mettre à disposition de l'utilisateur de multiples composants permettant de lire et d'alimenter divers types de sources et cibles de données et de générer en sortie un programme en code java interprétable par n'importe quel ordinateur. L'utilisateur met en place des « jobs » via l'interface graphique de Talend Open Studio sur lesquels il organise un processus à sa convenance.

Dans le cas de l'automatisation du contrôle qualité ProdInra, 5 jobs ont été développés :

- 1 Job Main (dit père) qui organise le lancement des 4 jobs suivants dits fils
- 1 Job Fils qui vérifie le format des fichiers reçus pour traitement
- 1 Job Fils qui traite les données du fichier d'export complet ProdInra
- 1 Job Fils qui traite les données du fichier d'export des auteurs ProdInra
- 1 Job Fils qui traite les données du fichier d'export des organismes externes ProdInra

[6] Laboratoire d'Etudes et de Recherches en Economie appliqué au Structures Marchés Agricoles, Ressources et Territoires, UMR Agrocampus Ouest-Inra
[7] Département de Recherche Inra en Sciences Sociales, Agriculture, Alimentation, Espace et Environnement (SAE2) dont fait partie l'UMR SMART-LERECO
[8] https://fr.talend.com/

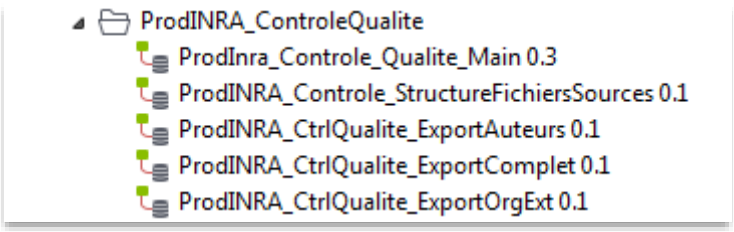

Figure 1. *Arborescence Talend Open Studio.*

Le logiciel Talend Open Studio a été utilisé car sa grande variété de composants permettant d'effectuer diverses opérations depuis une seule application. Enfin, une fois les jobs validés, l'outil offre la possibilité de générer un exécutable autonome déployable sur environnement Windows ou Linux.

Mode d'emploi d'installation de l'outil

Pré-requis

La version de Java installée sur le poste où sera déployé l'exécutable généré par Talend Open Studio doit être inférieure ou égale à celle qu'utilise Talend Open Studio lors de la génération de l'exécutable. La version actuelle de java utilisée dans notre cas est jre1.8.0_144.

Téléchargement

Pour récupérer l'exécutable, télécharger le zip à cette adresse : https://prodinra.inra.fr/record/447672

Installation

Afin de déployer en production l'exécutable de contrôle qualité ProdInra, il convient de procéder aux étapes suivantes.

- Dézipper le fichier ProdInra_Controle_Qualite_Main_*x.x*.zip (où x.x est le numéro de version livrée) sur votre poste. La décompression entraîne la création dans le répertoire « ProdInra_Controle_Qualite_Main_x.x » du contenu suivant :

Figure 2. *Contenu du répertoire ProdINRA_Controle_Qualite_Main_x.x*

- Le fichier Default.properties fichier permet de décrire le répertoire « sources » préalablement créé où sont stockés les fichiers à traiter (repFic_ExportProdInra) et le répertoire où devront être stockés les fichiers de résultats de contrôle (repFic_Controles). Pour pouvoir commencer, déclarer dans ce fichier le chemin à suivre pour les fichiers de sortie.

- Aller dans le répertoire « ProdInra_Controle_Qualite_Main_x.x\ProdInra_Controle_Qualite_Main\smart\prodinra_controle_qualite_main_x_x\contexts\ ». et modifier le chemin

Remarque : veiller à bien doubler les « \ » dans le chemin des répertoires à utiliser sinon cela ne fonctionnera pas (voir **Figure 3**).

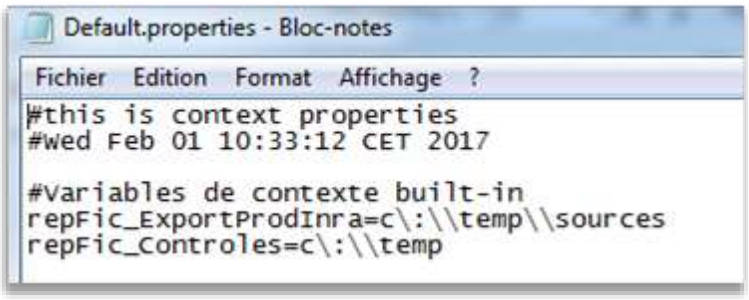

Figure 3. *Contenu du fichier Default.properties.*

Les différents tests pratiqués pour le contrôle de la qualité des données ProdInra nécessitent la présence d'un référentiel d'ISSN à comparer avec les ISSN présents dans le fichier des publications. Un fichier est présent par défaut, il contient les listes de revues en économie-gestion. Il peut facilement être complété ou remplacé par une autre liste si vous respectez la structure du fichier ISSN (deux colonnes : Titre revue / ISSN).

Dans ce cas, il convient de placer dans le même répertoire « sources » que les données extraites de ProdInra, un fichier nommé « Referentiel_ISSN.csv » constitué d'une seule colonne contenant l'ensemble des ISSN de référence à prendre en compte.

Exécution

Le programme à exécuter pour déclencher le contrôle des fichiers ProdInra est le fichier « **ProdInra_Controle_Qualite_Main_run.bat** » qui se trouve dans le répertoire « ProdInra_Controle_Qualite_Main_x.x\ProdInra_Controle_Qualite_Main ». Il génère un fichier zip.

Double-cliquer sur ce fichier .bat pour lancer le traitement.

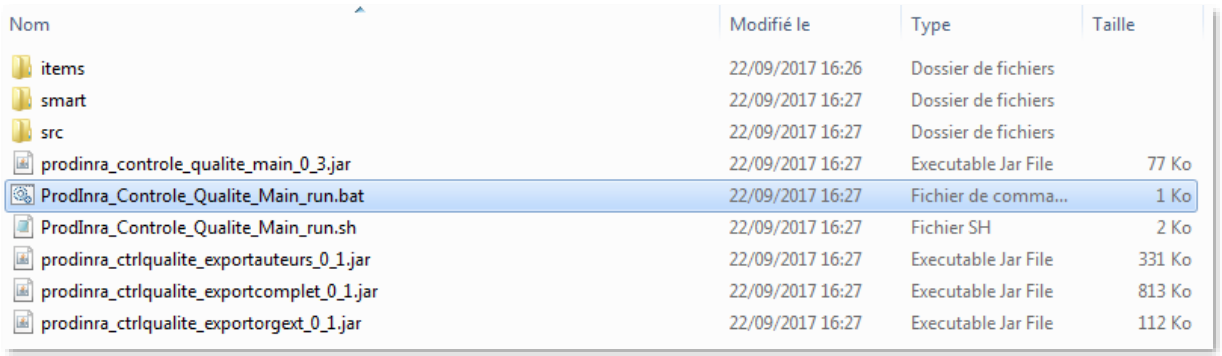

Figure 4*. Fichier à exécuter.*

L'exécutable analyse les différents fichiers présents dans le répertoire précisé dans la ligne repFic_ExportProdInra du fichier default.properties et génère si besoin en sortie des fichiers d'erreurs sur les différents tests pratiqués.

Le job mis en œuvre prévoit aussi la création d'un fichier de logs (généré au même endroit que le fichier zip final, dont le chemin a été déclaré dans Default.Properties). Celui-ci recense différentes informations permettant de comprendre les étapes réalisées lors du traitement : le répertoire indiqué dans le fichier de propriétés contenant les exports ProdInra à traiter, le nombre de fichiers à traiter, le nom des fichiers traités…

Figure 5. *Contenu du fichier de logs.*

Mode d'emploi lors de l'utilisation de l'exécutable

Fichiers sources à analyser

Les fichiers sources de l'outil doivent être extraits de ProdInra à partir d'une requête spécifique à vos besoins, effectuée en fonction du périmètre des données à analyser[9]. Il faut les déposer dans le dossier déclaré comme sources dans Default.Properties avec un format de nommage précis.

Il en faut trois pour que le script fonctionne correctement :

- Fichier export complet : ce fichier contient tous les champs associés aux notices dans ProdInra
  - Être nommé de façon à contenir la chaîne de caractères suivantes "exportcomplet.xls" (sans majuscules)

- Fichier Export bibliométrique éclaté par auteurs : ce fichier contient les informations d'affiliation Inra éclatées en détail par auteur

  - Être nommé de façon à contenir la chaîne de caractères suivante " exportauteur.xls" (sans majuscules)

  - Être nommé avec la même racine de nom que le fichier export complet (Exemple : PRD_20170127_SAE2_exportauteur.xls et PRD_20170127_SAE2_exportcomplet.xls ont la même racine "PRD_20170127_SAE2_")

- Fichier Export bibliométrique éclaté par organismes extérieurs : ce fichier contient les informations d'affiliation externes éclatées en détail pour chaque publication
  - Être nommé de façon à contenir la chaîne de caractères " exportorg.xls " (sans majuscules)
  - Être nommé avec la même racine de nom que le fichier export auteur (PRD_20170127_SAE2_exportorg.xls et PRD_20170127_SAE2_exportauteur.xls ont la même racine "PRD_20170127_SAE2_ »)

Tests effectués et précaution d'usages des fichiers de résultats

Cet outil n'est pas à utiliser tel quel. Il nécessite d'avoir une connaissance du corpus et de l'environnement scientifique pour lequel on analyse la qualité des données. Les tests réalisés à ce jour semblent être pertinents pour les échelles Unité ou Département. En effet, ces structures ont en commun des besoins pour des indicateurs sur le comité de lecture, le public cible, les auteurs (selon leur matricule). En plus des tests, nous avons automatisé la production de fichiers ou de listes (équivalentes à des tableaux croisés dynamiques) que nous produisions manuellement à chaque export pour gagner du temps.

[9] Par exemple, dans notre cas, on fait une requête dans ProdInra limitée aux 5 dernières années pour les publications dont l'affiliation contient SAE2.

Tableau 1. *Les Tests ou fichiers à analyser décrits avec les précautions d'usage*

Nom test ou fichier produit à analyser vous même	Contenu test	Précautions d'utilisation du fichier résultat
Test_Q1	Uniquement basé sur les articles à comité de lecture (ACL). La revue fait partie du « référentiel ISSN » déposée dans le dossier Source des fichiers à analyser et n'est pas déclarée ACL.	L'information est également à vérifier car une revue peut être non-scientifique avec comité de lecture. Ce test intéresse les documentalistes traitant des publications en Economie Gestion. Il est possible de changer la liste de validation pour mettre la liste de son choix (avec des ISSN) se créant son propre fichier ReferentielISSN.
Test_Q1b	Uniquement basé sur les ACL. L'article fait partie du Web of Science (WoS)[10] mais n'est pas déclaré ACL.	Vérifier si la revue est encore indexée dans le WoS / Ne résout pas le problème des revues nouvelles entrantes dans le WoS.
Test_Q2	Recherche les auteurs qui ont déjà porté une fois un matricule dans le corpus et identifie les notices où ils n'en portent pas. Il arrive que le rattachement à une UMR Inra ne soit pas clair dans la publication ou que le relecteur ne connaisse pas cet auteur ou inversement que les personnes aient été Inra et soient encore rattachées à une entité après leur départ.	Il faut vérifier dans la base RH (voir avec l'Unité ou le Département) si la personne était présente (ou non) à cette date. Il est nécessaire d'avoir un contact avec les agents RH pour mener à bien ce travail.
Test_Q3	Uniquement basé sur les articles. La revue a un identifiant ISSN renseigné, mais l'article n'a pas de clé UT (identifiant WoS), contrairement à d'autres articles de la même revue présents dans ProdInra.	Vérifier si la revue est encore indexée dans le WoS ou si l'article est récent, dans ce cas il peut être normal de ne pas avoir de clé UT.
Test_Q4	Présence d'une donnée dans le champ unité expérimentale	Vérifier si la publication est écrite avec un co-auteur affilié à un autre Département, à une Unité expérimentale ou s'il existe des partenariats avec une Unité expérimentale (souvent spécifié dans la publication).
Test_Q5	La revue de l'article étudié n'a pas d'identifiant ISSN renseigné.	Vérifier sur le site de la revue le numéro ISSN (imprimé en priorité, électronique pour les revues exclusivement en ligne).
ConfInvitee	Liste les conférences invitées pour vérifier si elles en sont bien	Ce fichier est à faire vérifier par les chercheurs eux-mêmes. Définition d'une conférence invitée à leur fournir : « communication présentée lors d'une conférence scientifique ou non où l'agent est le keynote speaker »
TitreConf	Liste les intitulés des conférences dans le but de les uniformiser : nécessite ensuite la production d'un tableau croisé dynamique pour identifier les doublons	Le référentiel des noms de colloques n'étant pas un référentiel contrôlé dans ProdInra, le résultat de ce fichier peut être conséquent. Il est important de faire ce chantier qualité en concertation avec l'équipe ProdInra et ainsi bénéficier des possibilités de correction par lots.

[10] Définition WoS : Web of Science. Base de données bibliographiques internationale pluridisciplinaire produite par Clarivate Analytics. Accès LDAP.

Nom test ou fichier produit à analyser vous même	Contenu test	Précautions d'utilisation du fichier résultat
These	École doctorale / Laboratoire d'accueil / Établissement de soutenance	Vérifier que les 3 champs soient remplis avec des données cohérentes et justes. Usage : en déduire un référentiel de saisie à diffuser auprès des professionnels de l'IST concernés par le corpus étudié, après correction.
OrgExt	Dans ce fichier on liste des notices dont les éléments qui ont été rentrés comme institution alors qu'ils pourraient être des subdivisions (ex : l'Inra est une institution / SMART-LERECO est une subdivision)	Vérifier au cas par cas manuellement.

Fichiers générés par l'exécutable de contrôle de qualité

Après un clic sur le fichier .bat voici les tests logiques effectués par le script. Le langage n'est pas forcément naturel à la lecture pour un non informaticien. Un accompagnement personnalisé et/ou un guide de lecture joint au script est envisageable.

Les fichiers générés par le programme sont contenus dans un fichier zip qui apparaîtra à l'emplacement de l'ordinateur indiqué dans le fichier default.properties.

Les tests réalisés sur le fichier d'export complet produisent en sortie les fichiers suivants :

- Le fichier *Nomfichiertraite*Publi_Out_final.xls qui recense pour l'ensemble des publications présentes dans le fichier d'export complet le résultat (OK ou KO) aux différents tests
- Le fichier *Nomfichiertraite*TCD_These_Out.xls contient toutes les publications de type « THESIS » avec les informations relatives à l'école doctorale associée
- N*omfichiertraite*TCD_TitreConf liste l'ensemble des valeurs de la colonne « Evenement »
- N*omfichiertraite*_Q2_Ko_TypeProduit liste l'ensemble des publications ayant échoué au test Q2
- N*omfichiertraite*_Q4_Ko liste l'ensemble des publications ayant échoué au test Q4

Les tests réalisés sur le fichier des auteurs éclatés produisent en sortie les fichiers suivants :

- Le fichier *Nomfichiertraite*_Auteurs_out_final.xls qui contient pour l'ensemble des publications du fichier export auteurs les résultats aux tests mis en œuvre
- Le fichier *Nomfichiertraite*Auteurs_Q2_KO_out.xls contient les publications dont l'auteur mentionné présente un matricule vide alors que ce même auteur dispose d'un matricule sur d'autres publications du même fichier
- Le fichier *Nomficihiertraite*TCD_ConfInvitee_out.xls qui contient pour chaque Unité les publications mentionnant une conférence invitée
- Le fichier *Nomficihiertraite*_Q2_Ko.xls qui contient les publications ayant échoué au test Q2

Les tests réalisés sur le fichier d'export des organismes produisent en sortie les fichiers suivants :

- Le fichier *Nomfichiertraite*OrgExt_final.xls contient les publications issues du fichier d'export bibliométrique éclaté par organisme externe avec la mention indiquant si le contenu du champ Nom affiliation Externe contient les mots « unit », « labo », « centre » ou « groupe ». Si ce champ contient les mots recherchés, le résultat du test est KO.

Conclusions et perspectives

Une fois que vous avez ces fichiers et en suivant bien les précautions d'usages listées dans le **Tableau 1**, les modifications dans ProdInra sont effectuées au cas par cas ou en concertation avec l'équipe ProdInra pour celles qui peuvent faire l'objet d'une modification par lot.

Les documentalistes du Département SAE2 passaient beaucoup de temps à reproduire très régulièrement les mêmes fichiers pour vérifier la qualité des données. Grâce à cet outil simple (trois exports ProdInra puis un clic pour produire les fichiers), on peut espacer les traitements qualité sans perdre en efficacité.

Quand on se place du point de vue d'un informaticien, cet outil est simple à développer et facile à mettre en œuvre. Du point de vue des personnels IST, c'est un immense gain de temps.

Ce script est souple. Nous avons donc la possibilité d'ajouter des tests après analyse de la faisabilité par nos soins. Pour cela, il suffit de nous contacter.

ProdInra est voué à terme à migrer vers HaL, le script peut tout à fait être réadapté aux formats d'export proposés par HaL. Il y aura dans ce cas un travail d'évolution du script à prévoir.

Par ailleurs, nous réfléchissons à développer un package dans R[11] qui embarquerait les fonctionnalités décrites ici ainsi que des outils de production d'indicateurs à destination des Unités du Département SAE2[12].

Les auteurs remercient Agnès Girard, Diane Le Henaff et Marie-Violaine Tatry pour la relecture et les propositions d'évolution de cet article.

[11] R est un langage de programmation et un logiciel libre dédié aux statistiques et à la science des données. Voir Wikipédia, consulté le 9/02/2018.

[12] Cariou, S., Frappier, L. (2018). Bibliométrie, Contrôle qualité et production d'indicateurs. Automatisation sous R. présenté à 7. Rencontres R, Rennes, FRA (2018-07-04 - 2018-07-06). https://prodinra.inra.fr/record/435029