

# Application de Data Cleansing

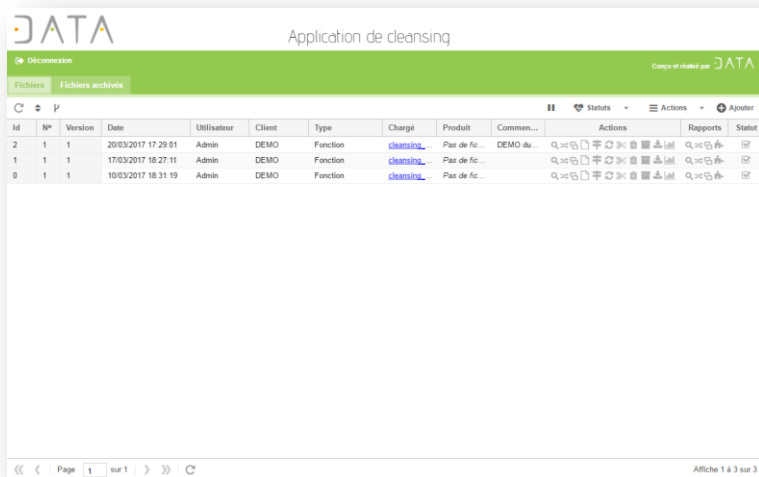
Mise en œuvre du processus de  
valorisation de la qualité des  
données

Avril 2017

# Un contexte unique à chaque entreprise

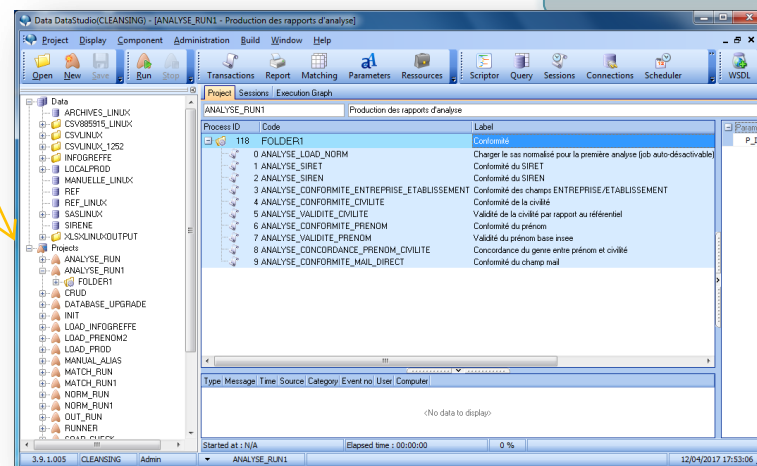
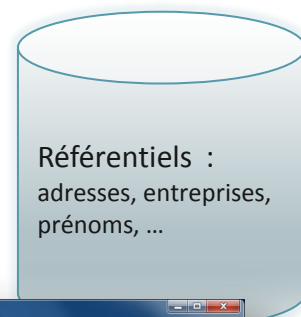
- Chaque entreprise a ses spécificités au niveau de ses données.
- L'application Data Cleansing met en œuvre des modules personnalisés à votre contexte.
  - Ces modules sont groupés par étapes :
    - Vérification de la forme de la donnée (Conformité à un format),
    - Validité d'une donnée par rapport à un référentiel ou une règle de validation (ex : la clé du RIB),
    - Concordance de deux données dépendantes entre-elles (exemple : le genre et le prénom),
    - Déduplication et Matching par rapport aux données de référence

# Architecture modulaire



Interface web ergonomique :

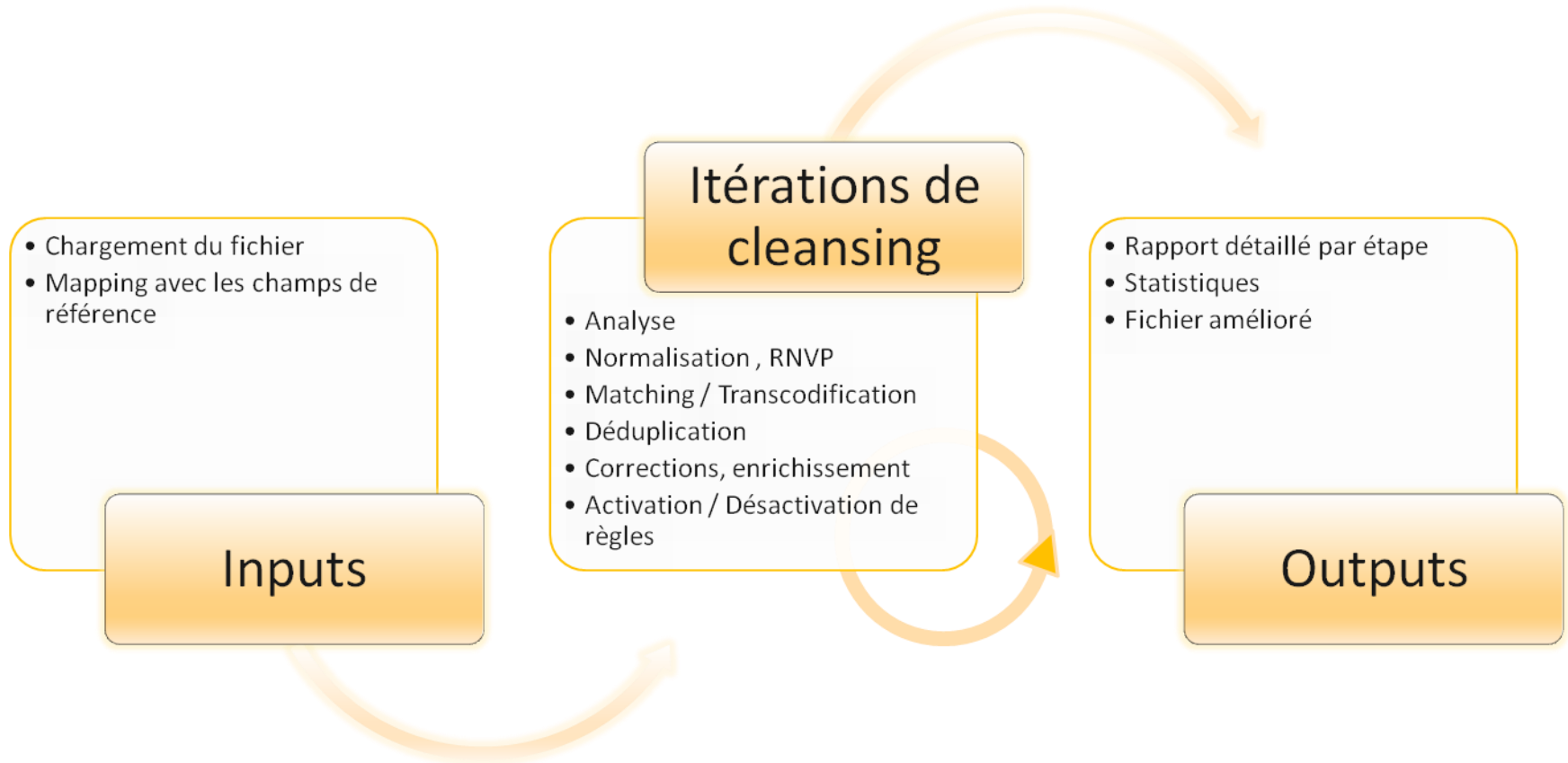
- Login
- Gestion des sources de données à traiter
- Gestion des modules de data cleansing
- Lancement des itérations de data cleansing
- Reporting / Statistiques



Serveur DataStudio :

- Paramétrage des modules de qualité
- Exécution des traitements
- Gestion des accès

# Processus de cleansing



# L'interface de gestion

Application de cleansing

Déconnexion

Conçu et réalisé par DATA

Fichiers Fichiers archivés

Id	N°	Version	Date	Utilisateur	Client	Type	Chargé	Produit	Commentaires	Actions	Rapports	Statut
2	1	1	20/03/2017 17:29:01	Admin	DEMO	Fonction	cleansing...	Pas de fic...	DEMO du...	[Icons]	[Icons]	[Check]
1	1	1	17/03/2017 18:27:11	Admin	DEMO	Fonction	cleansing...	Pas de fic...		[Icons]	[Icons]	[Check]
0	1	1	10/03/2017 18:31:19	Admin	DEMO	Fonction	cleansing...	Pas de fic...		[Icons]	[Icons]	[Check]

Mise à jour des données de référence

Accès au fichier à « nettoyer »

Étapes du processus de cleansing

Rapport détaillé ligne par ligne de chaque étape du processus

Page 1 sur 1

Affiche 1 à 3 sur 3

Analyse

Normalisation , RNVP

Matching / Transco.

Déduplication

# Mapping des données sources

Mapping
✕

Options de parsing

Jeu de caractères :

Caractère de séparation :

Ligne d'en-têtes :

Nombre maximal de lignes à afficher :

Nombre maximal de colonnes :

Fin de ligne :

Caractère de délimitation des chaînes :

Première ligne de données :

Taille maximale à pré-traiter (Mo) :

Taille maximale d'une colonne :

Prévisualisation ↻

ENTREPRI...	ISIN	TRADING LOCATION
ACCOR	FR0000120...	Euronext Paris
AIR LIQUIDE	FR0000120...	Euronext Paris
AIRBUS	NL0000235...	Euronext Paris
ARCELOR...	LU0323134...	Euronext Amsterdam
ATOS	FR0000051...	Euronext Paris
AXA	FR0000120...	Euronext Paris

Mapping des champs 🗑️ ✓

Champ fichier	Champ mappé
ENTREPRISE	ENTREPRISE
ISIN	
TRADING LOC...	

Structure du fichier à traiter

Visualisation des données sources

Mapping des champs du fichier avec les colonnes de référence

# Analyse

Lancement et production du rapport d'exécution

DATA Application de cleansing

↳ Déconnexion Conçu et réalisé par DATA

Fichiers **Operations - cleansing\_demo\_2.csv** Fichiers archivés

🔄 Lancer ✓ Sauvegarder

Projet	Folder	Job	Actif
Production des rapports d'analyse	Conformité	Charger le sas normalisé pour la première analys...	<input type="checkbox"/>
Production des rapports d'analyse	Conformité	Conformité du SIRET	<input type="checkbox"/>
Production des rapports d'analyse	Conformité	Conformité du SIREN	<input type="checkbox"/>
Production des rapports d'analyse	Conformité	Conformité des champs ENTREPRISE/ETABLISS...	<input type="checkbox"/>
Production des rapports d'analyse	Conformité	Conformité de la civilité	<input checked="" type="checkbox"/>
Production des rapports d'analyse	Conformité	Validité de la civilité par rapport au référentiel	<input checked="" type="checkbox"/>
Production des rapports d'analyse	Conformité	Conformité du prénom	<input checked="" type="checkbox"/>
Production des rapports d'analyse	Conformité	Validité du prénom base insee	<input type="checkbox"/>
Production des rapports d'analyse	Conformité	Concordance du genre entre prénom et civilité	<input checked="" type="checkbox"/>
Production des rapports d'analyse	Conformité	Conformité du champ mail	<input checked="" type="checkbox"/>

Liste des règles d'Analyse de la conformité (Extensible, règles personnalisables)

Sélection des règles à appliquer

# Rapport d'analyse

Données - RAPPORT\_ANALYSE

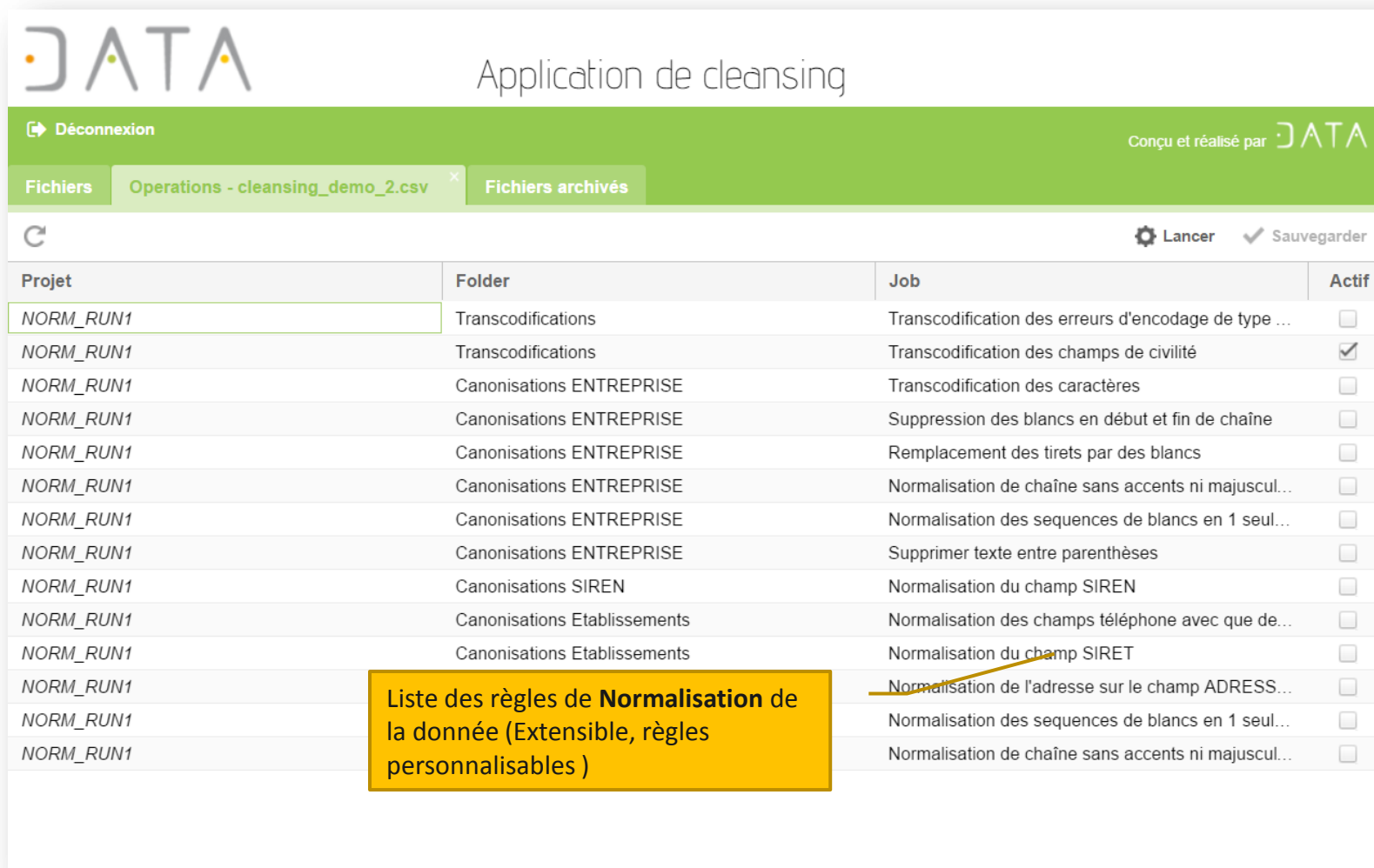
Page 1 sur 2 | Taille des pages : 25 | Affiche 1 à 25 sur 26

LIGNE	DOSSIER	JOB	LIBELLE_ERREUR
3	Conformité	Concordance du genre entre prénom et civilité	La valeur du champ PRENOM="Roger" et du champ CIVILITE="Madame" , ne concordent pas avec le genre
5	Conformité	Conformité du champ mail	La valeur du champ MAIL_DIRECT="olivier.bouas-laurent@@generationmedia.fr" n'est pas de la forme xxx@
9	Conformité	Conformité du prénom	La valeur du champ PRENOM="FrÃ©dÃ©ric" n'est pas alphanumérique
23	Conformité	Conformité du champ mail	La valeur du champ MAIL_DIRECT="" n'est pas de la forme xxx@domain
31	Conformité	Conformité du champ mail	La valeur du champ MAIL_DIRECT="" n'est pas de la forme xxx@domain
33	Conformité	Conformité du champ mail	La valeur du champ MAIL_DIRECT="" n'est pas de la forme xxx@domain
52	Conformité	Conformité du champ mail	La valeur du champ MAIL_DIRECT="" n'est pas de la forme xxx@domain
62	Conformité	Conformité du champ mail	La valeur du champ MAIL_DIRECT="" n'est pas de la forme xxx@domain
72	Conformité	Conformité du champ mail	La valeur du champ MAIL_DIRECT="" n'est pas de la forme xxx@domain
91	Conformité	Conformité du champ mail	La valeur du champ MAIL_DIRECT="" n'est pas de la forme xxx@domain
99	Conformité	Conformité du champ mail	La valeur du champ MAIL_DIRECT="" n'est pas de la forme xxx@domain
136	Conformité	Conformité du champ mail	La valeur du champ MAIL_DIRECT="" n'est pas de la forme xxx@domain
163	Conformité	Conformité du champ mail	La valeur du champ MAIL_DIRECT="" n'est pas de la forme xxx@domain
237	Conformité	Conformité du champ mail	La valeur du champ MAIL_DIRECT="" n'est pas de la forme xxx@domain
287	Conformité	Conformité du champ mail	La valeur du champ MAIL_DIRECT="" n'est pas de la forme xxx@domain
292	Conformité	Conformité du champ mail	La valeur du champ MAIL_DIRECT="" n'est pas de la forme xxx@domain
296	Conformité	Conformité du champ mail	La valeur du champ MAIL_DIRECT="" n'est pas de la forme xxx@domain
309	Conformité	Conformité du champ mail	La valeur du champ MAIL_DIRECT="" n'est pas de la forme xxx@domain
327	Conformité	Conformité du champ mail	La valeur du champ MAIL_DIRECT="" n'est pas de la forme xxx@domain

Description du problème de conformité par cellule (colonne et ligne)



# Normalisation



The screenshot shows the 'Application de cleansing' interface. At the top, there is a green navigation bar with 'Déconnexion' on the left and 'Conçu et réalisé par DATA' on the right. Below this, there are tabs for 'Fichiers', 'Operations - cleansing\_demo\_2.csv', and 'Fichiers archivés'. A toolbar contains a refresh icon, a gear icon labeled 'Lancer', and a checkmark icon labeled 'Sauvegarder'. The main content is a table with the following columns: 'Projet', 'Folder', 'Job', and 'Actif'. The table lists 15 rows of normalization rules for the project 'NORM\_RUN1'. A yellow callout box highlights the last three rows of the table.

Projet	Folder	Job	Actif
NORM_RUN1	Transcodifications	Transcodification des erreurs d'encodage de type ...	<input type="checkbox"/>
NORM_RUN1	Transcodifications	Transcodification des champs de civilité	<input checked="" type="checkbox"/>
NORM_RUN1	Canonisations ENTREPRISE	Transcodification des caractères	<input type="checkbox"/>
NORM_RUN1	Canonisations ENTREPRISE	Suppression des blancs en début et fin de chaîne	<input type="checkbox"/>
NORM_RUN1	Canonisations ENTREPRISE	Remplacement des tirets par des blancs	<input type="checkbox"/>
NORM_RUN1	Canonisations ENTREPRISE	Normalisation de chaîne sans accents ni majuscul...	<input type="checkbox"/>
NORM_RUN1	Canonisations ENTREPRISE	Normalisation des sequences de blancs en 1 seul...	<input type="checkbox"/>
NORM_RUN1	Canonisations ENTREPRISE	Supprimer texte entre parenthèses	<input type="checkbox"/>
NORM_RUN1	Canonisations SIREN	Normalisation du champ SIREN	<input type="checkbox"/>
NORM_RUN1	Canonisations Etablissements	Normalisation des champs téléphone avec que de...	<input type="checkbox"/>
NORM_RUN1	Canonisations Etablissements	Normalisation du champ SIRET	<input type="checkbox"/>
NORM_RUN1		Normalisation de l'adresse sur le champ ADRESS...	<input type="checkbox"/>
NORM_RUN1		Normalisation des sequences de blancs en 1 seul...	<input type="checkbox"/>
NORM_RUN1		Normalisation de chaîne sans accents ni majuscul...	<input type="checkbox"/>

Liste des règles de **Normalisation** de la donnée (Extensible, règles personnalisables )

# Rapport de normalisation

Données - RAPPORT\_NORMALISE

Page 1 sur 20 Taille des pages : 25 Affiche 1 à 25 sur 500

WRK_ID_ENREG	ANY_ISUPD	CIVILITE_ISUPD	CIVILITE	N_CIVILITE	NOM_ISUPD	PRENOM_ISUPD	PRENOM	N_PRENOM	DATE_NAISSANCE_ISUPD
1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	M.	Monsieur	<input type="checkbox"/>	<input type="checkbox"/>	Pierre	Pierre	<input type="checkbox"/>
2	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Mme	Madame	<input type="checkbox"/>	<input type="checkbox"/>	Frédérique	Frédérique	<input type="checkbox"/>
3	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Mme	Madame	<input type="checkbox"/>	<input type="checkbox"/>	Roger	Roger	<input type="checkbox"/>
4	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	M.	Monsieur	<input type="checkbox"/>	<input type="checkbox"/>	Xavié	Xavié	<input type="checkbox"/>
5	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	M.	Monsieur	<input type="checkbox"/>	<input type="checkbox"/>	Olivier	Olivier	<input type="checkbox"/>
6	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	M.	Monsieur	<input type="checkbox"/>	<input type="checkbox"/>	Jean-Luc	Jean-Luc	<input type="checkbox"/>
7	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	M.	Monsieur	<input type="checkbox"/>	<input type="checkbox"/>	Laurent	Laurent	<input type="checkbox"/>
8	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	M.	Monsieur	<input type="checkbox"/>	<input type="checkbox"/>	Michel	Michel	<input type="checkbox"/>
9	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	M.	Monsieur	<input type="checkbox"/>	<input type="checkbox"/>	FrÃ©dÃ©ric	FrÃ©dÃ©ric	<input type="checkbox"/>
10	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	M.	Monsieur	<input type="checkbox"/>	<input type="checkbox"/>	Philippe	Philippe	<input type="checkbox"/>
11	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Mme	Madame	<input type="checkbox"/>	<input type="checkbox"/>	Sylvie	Sylvie	<input type="checkbox"/>
12	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	M.	Monsieur	<input type="checkbox"/>	<input type="checkbox"/>	Olivier	Olivier	<input type="checkbox"/>
13	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	M.	Monsieur	<input type="checkbox"/>	<input type="checkbox"/>	Robert	Robert	<input type="checkbox"/>
14	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	M.	Monsieur	<input type="checkbox"/>	<input type="checkbox"/>	Claude	Claude	<input type="checkbox"/>
15	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	M.	Monsieur	<input type="checkbox"/>	<input type="checkbox"/>	Bertrand	Bertrand	<input type="checkbox"/>
16	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	M.	Monsieur	<input type="checkbox"/>	<input type="checkbox"/>	Stephan	Stephan	<input type="checkbox"/>
17	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	M.	Monsieur	<input type="checkbox"/>	<input type="checkbox"/>	Eric	Eric	<input type="checkbox"/>
18	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	M.	Monsieur	<input type="checkbox"/>	<input type="checkbox"/>	Georges	Georges	<input type="checkbox"/>
19	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	M.	Monsieur	<input type="checkbox"/>	<input type="checkbox"/>	Christian	Christian	<input type="checkbox"/>

Rapport avant et après normalisation

Possibilité d'action (exemple correction du charset ou compléter la transcodification) pour prise en compte à la nouvelle **itération** de normalisation

# Matching

Application de cleansing

DATA

Déconnexion

Conçu et réalisé par DATA

Fichiers Operations - cleansing\_demo\_2.csv Fichiers archivés

Filtre de matching : Type de matching : Indépendant Lancer Sauvegarder

Projet	Folder	Job	Actif
MATCH_RUN1	Matches simples	MATCH_SIRET	<input type="checkbox"/>
MATCH_RUN1	Matches simples	MATCH_SIREN	<input type="checkbox"/>
MATCH_RUN1	Matches simples	MATCH_ALIAS_EXTERNE	<input type="checkbox"/>
MATCH_RUN1	Matches simples	MATCH_ALIAS_NOMI	<input type="checkbox"/>
MATCH_RUN1	Matches simples	Match des SIREN sur les alias de siren multisiren ...	<input type="checkbox"/>
MATCH_RUN1	Matches simples	Match des noms d'entreprise sur les alias de nom...	<input type="checkbox"/>
MATCH_RUN1	Matches simples	MATCH_ENTREPRISE_DOMAINE_WEBSITE	<input type="checkbox"/>
MATCH_RUN1	Matches simples	MATCH_ENTREPRISE_DOMAINE_MAIL	<input type="checkbox"/>
MATCH_RUN1	Matches simples	MATCH_ENTREPRISE_DOMAINE_MAILDIRECT	<input type="checkbox"/>
MATCH_RUN1	Matches simples	MATCH_FUZZY	<input type="checkbox"/>
MATCH_RUN1	Matches simples	Matching sur Civilité-Prénom (Fuzzy)-Nom	<input checked="" type="checkbox"/>

Liste des règles de **Matching** par rapport aux données de référence (Extensible, règles personnalisables )

Matching exact, flou (fuzzy) ou autre algorithme à définir

# Rapport de Matching

Données - RAPPORT\_MATCHING \*

Page 1 sur 1 | Taille des pages : 50 | Affiche 1 à 40 sur 40

O_MANUAL	LIGNE ↑	ENTREPRISE	O_NUM	O_DOUBLON	O_STATUT	O_ENTREPRISE	O_SIREN	O_WEBSITE
<input type="checkbox"/>	2	AIR LIQUIDE	0	0	MATCHMUL	Air Liquide	450656764	http://www.airliquide.com
<input type="checkbox"/>	2	AIR LIQUIDE	1	0	MATCHMUL	Air Liquide	552096281	http://www.airliquide.com
<input type="checkbox"/>	3	AIRBUS	0	0	MATCHMUL	Airbus	341535094	http://www.airbusgroup.com
<input type="checkbox"/>	3	AIRBUS	1	0	MATCHMUL	Airbus	383474814	http://www.airbus.com
<input type="checkbox"/>	4	ARCELORMITTAL	0	0	MATCHMUL	ArcelorMittal	562094425	http://www.arcelormittal.com
<input type="checkbox"/>	4	ARCELORMITTAL	1	0	MATCHMUL	ArcelorMittal	336880463	http://www.pum.fr
<input type="checkbox"/>	5	ATOS	0	0	MATCHMUL	Atos	323623603	http://fr.atos.net
<input type="checkbox"/>	8	BOUYGUES	0	0	MATCHMUL	Bouygues	324295757	
<input type="checkbox"/>	8	BOUYGUES	1	0	MATCHMUL	Bouygues	572015246	http://www.bouygues.fr
<input type="checkbox"/>	11	CREDIT AGRICOLE	0	0	MATCHMUL	Credit agricole	314886797	http://www.ca-nord-est.fr
<input type="checkbox"/>	11	CREDIT AGRICOLE	1	0	MATCHMUL	Crédit agricole	784608416	http://www.credit-agricole.com
<input type="checkbox"/>	12	DANONE	0	0	MATCHMUL	Danone	552032534	http://www.danone.com
<input type="checkbox"/>	13	ENGIE	0	0	MATCHMUL	Engie	542107651	http://www.engie.com
<input type="checkbox"/>	15	KERING	0	0	MATCHMUL	Kering	552075020	http://www.kering.com
<input type="checkbox"/>	16	L'OREAL	0	0	MATCHMUL	L'Oréal	632012100	http://www.loreal.com
<input type="checkbox"/>	18	LEGRAND	0	0	MATCH	Legrand	758501001	http://www.legrand.com
<input type="checkbox"/>			0	0	MATCH	LVMH	775670417	http://www.lvmh.com
<input type="checkbox"/>			0	0	MATCH	Nokia		
<input type="checkbox"/>			0	0	MATCHMUL	Orange	380129866	http://www.orange.com

Données provenant du référentiel après matching

Enregistrer

Merci de votre attention

Retrouvez-nous sur [www.data.fr](http://www.data.fr)