Application de Data Cleansing

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

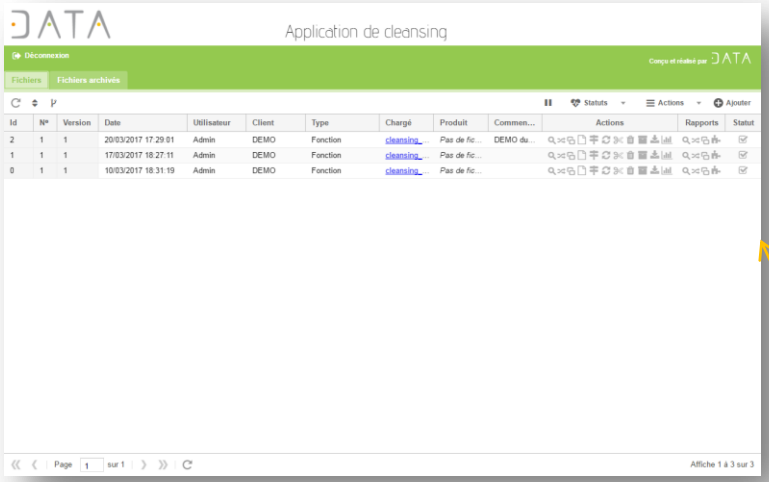
Avril 2017

© Copyright 2017 DATA

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

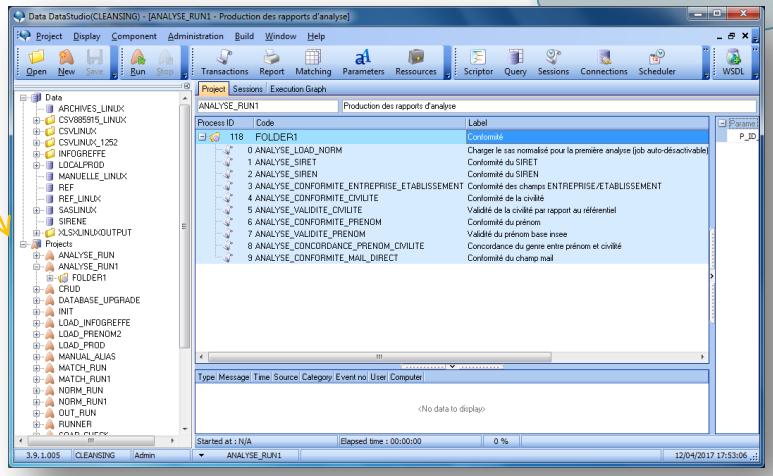
Architecture modulaire



Interface web ergonomique :

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

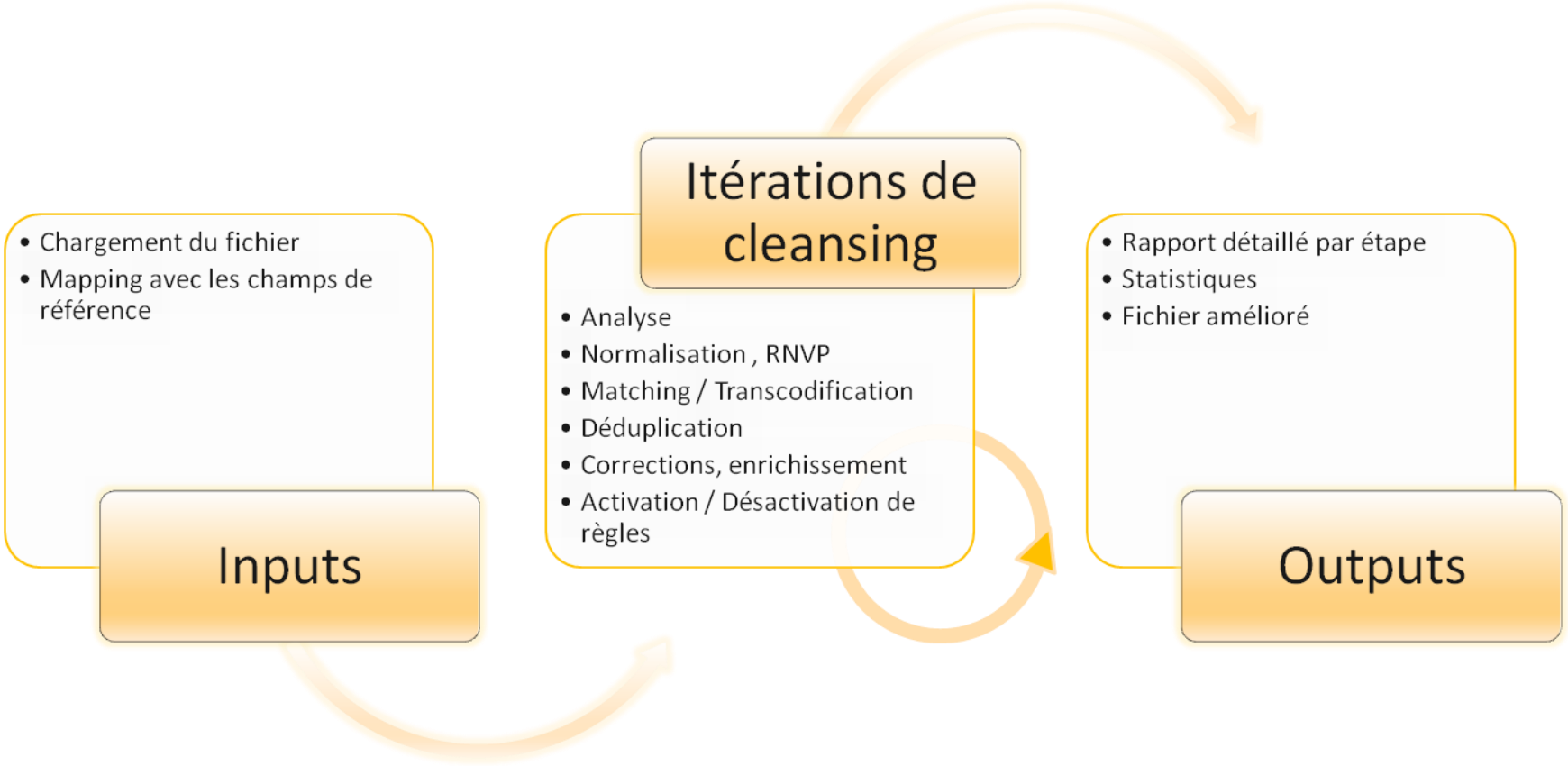
Référentiels :
adresses, entreprises, prénoms, …



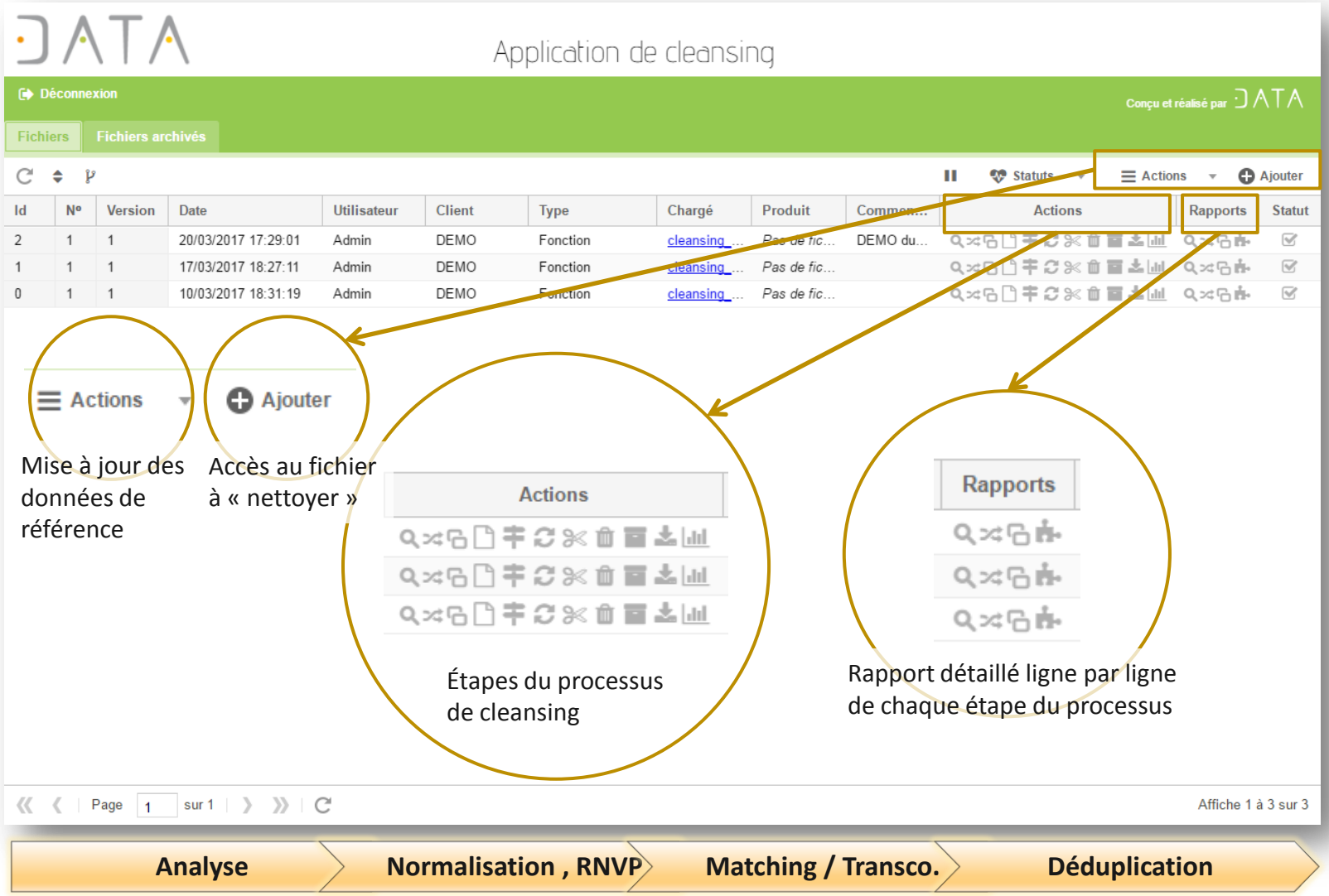
Serveur DataStudio :

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

Processus de cleansing

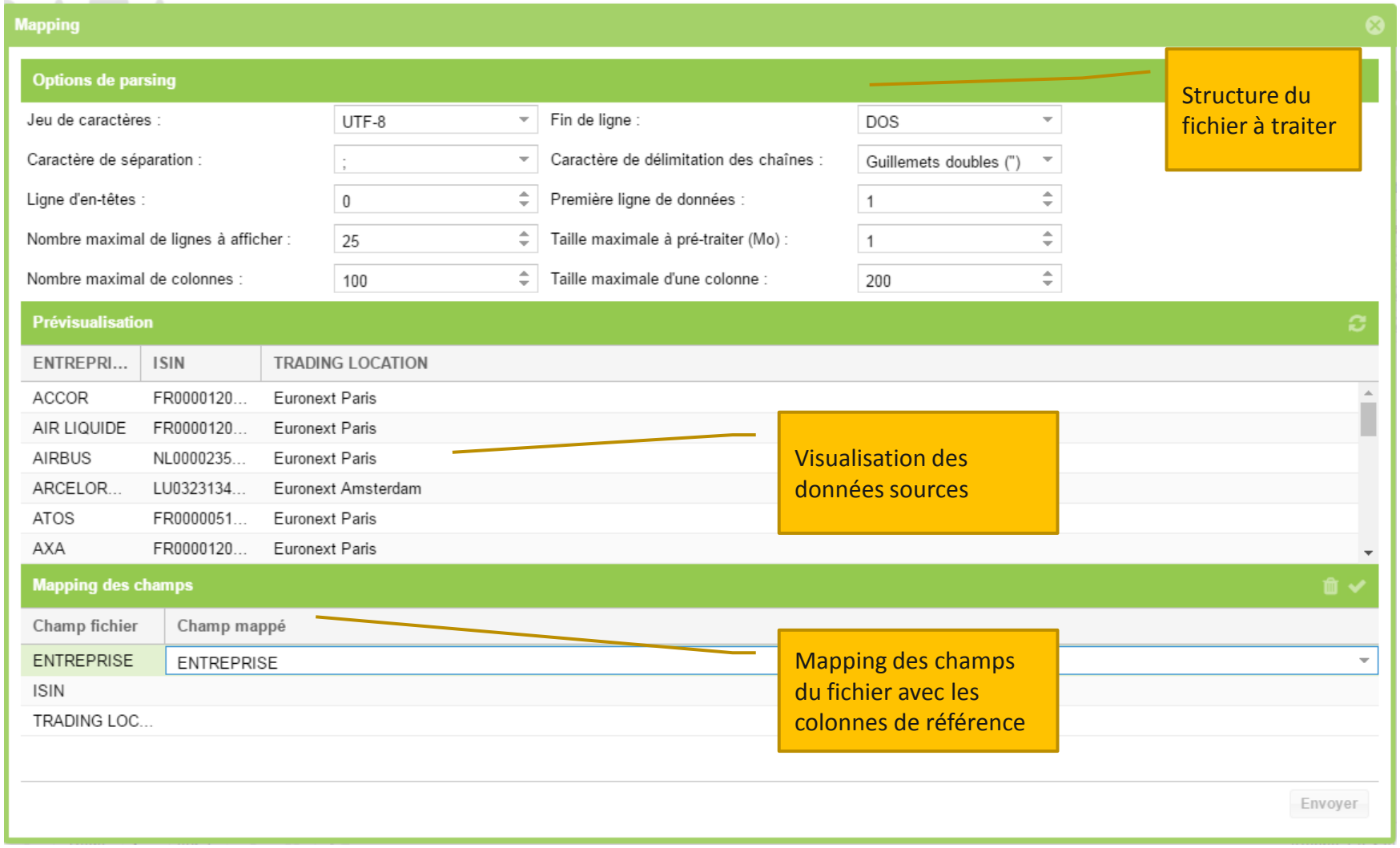


L'interface de gestion



Analyse > **Normalisation , RNVP** > **Matching / Transco.** > **Déduplication**

Mapping des données sources



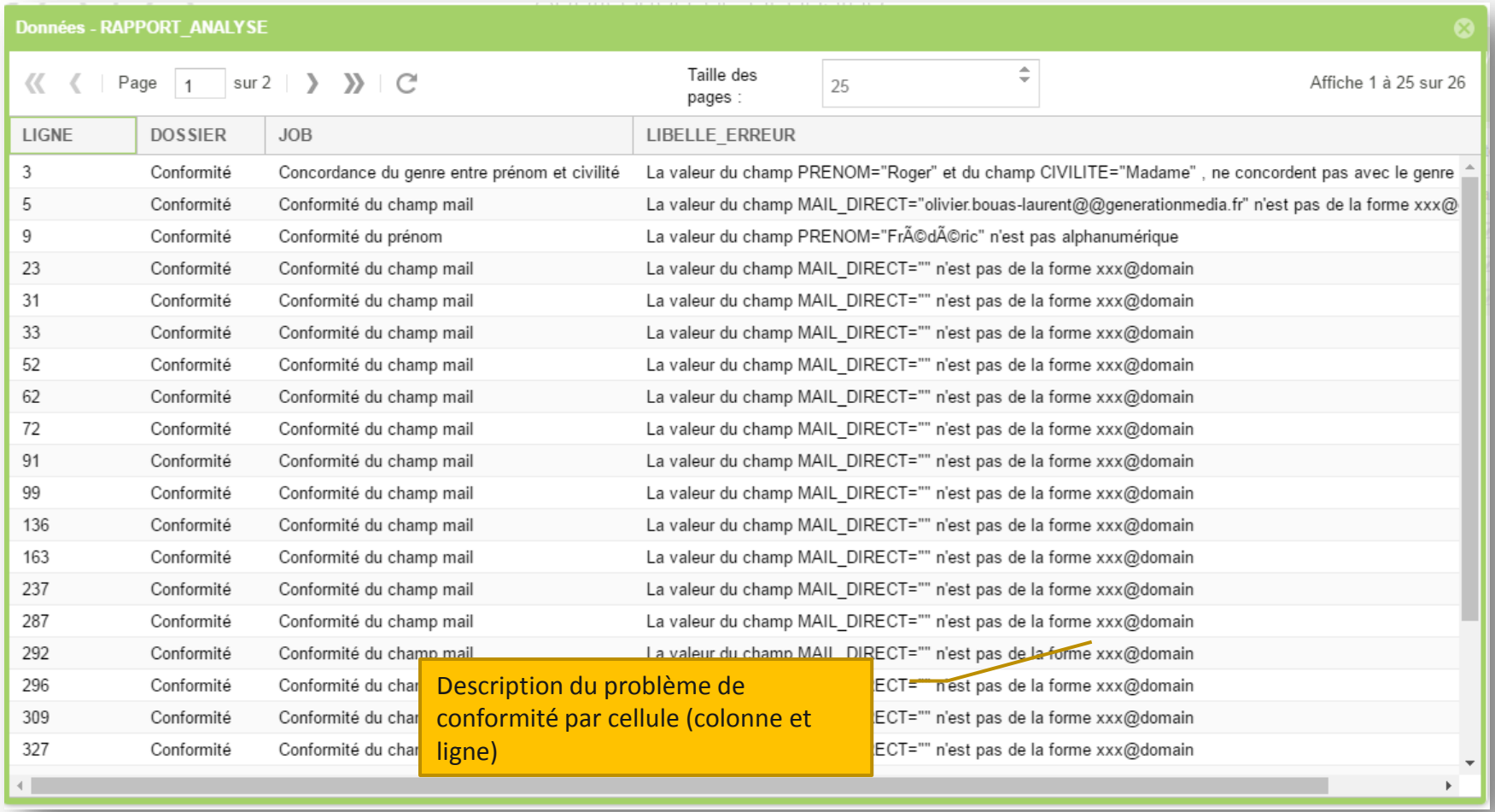
Analyse

Lancement et production du rapport d'exécution

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

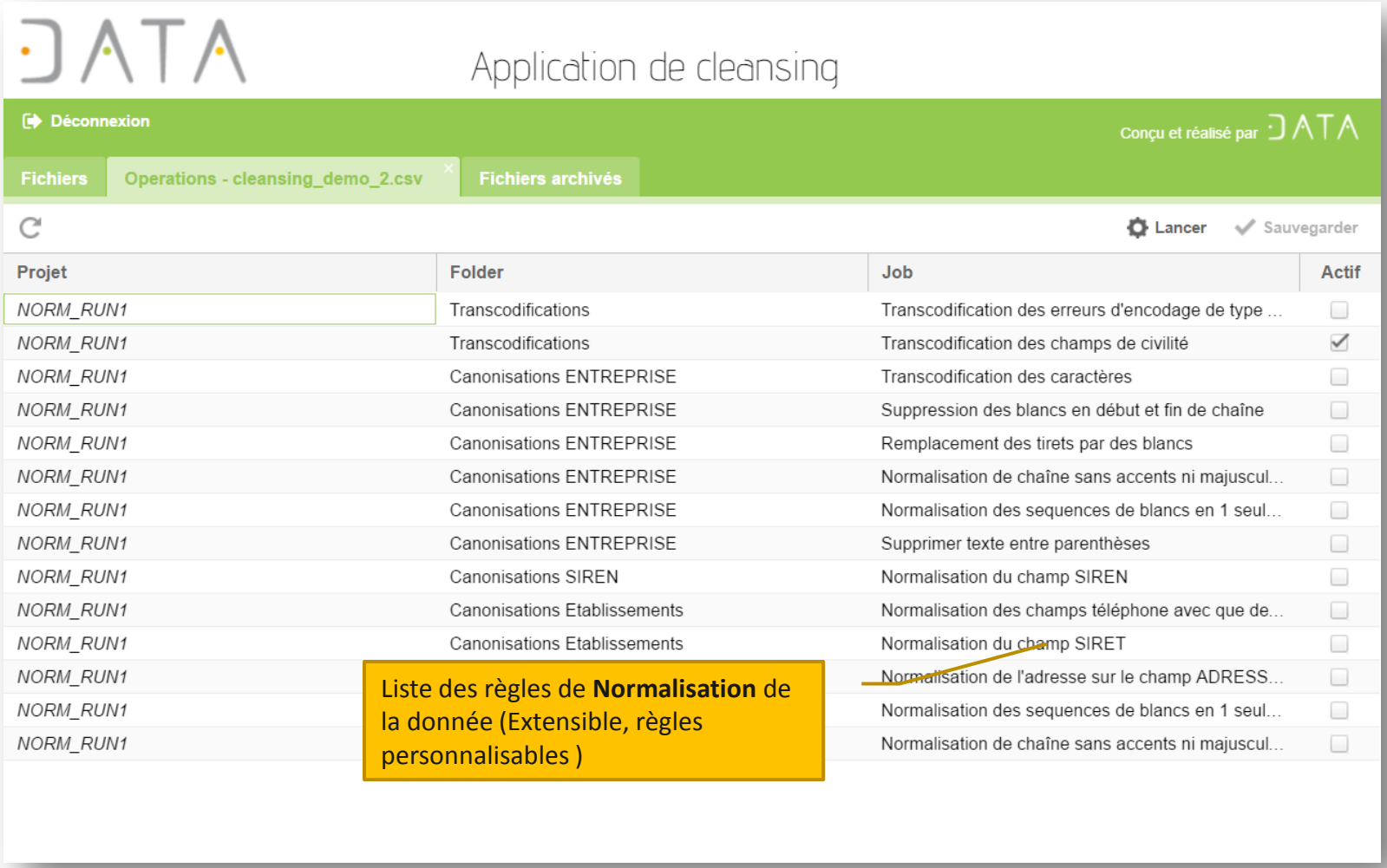
Sélection des règles à appliquer

Rapport d'analyse

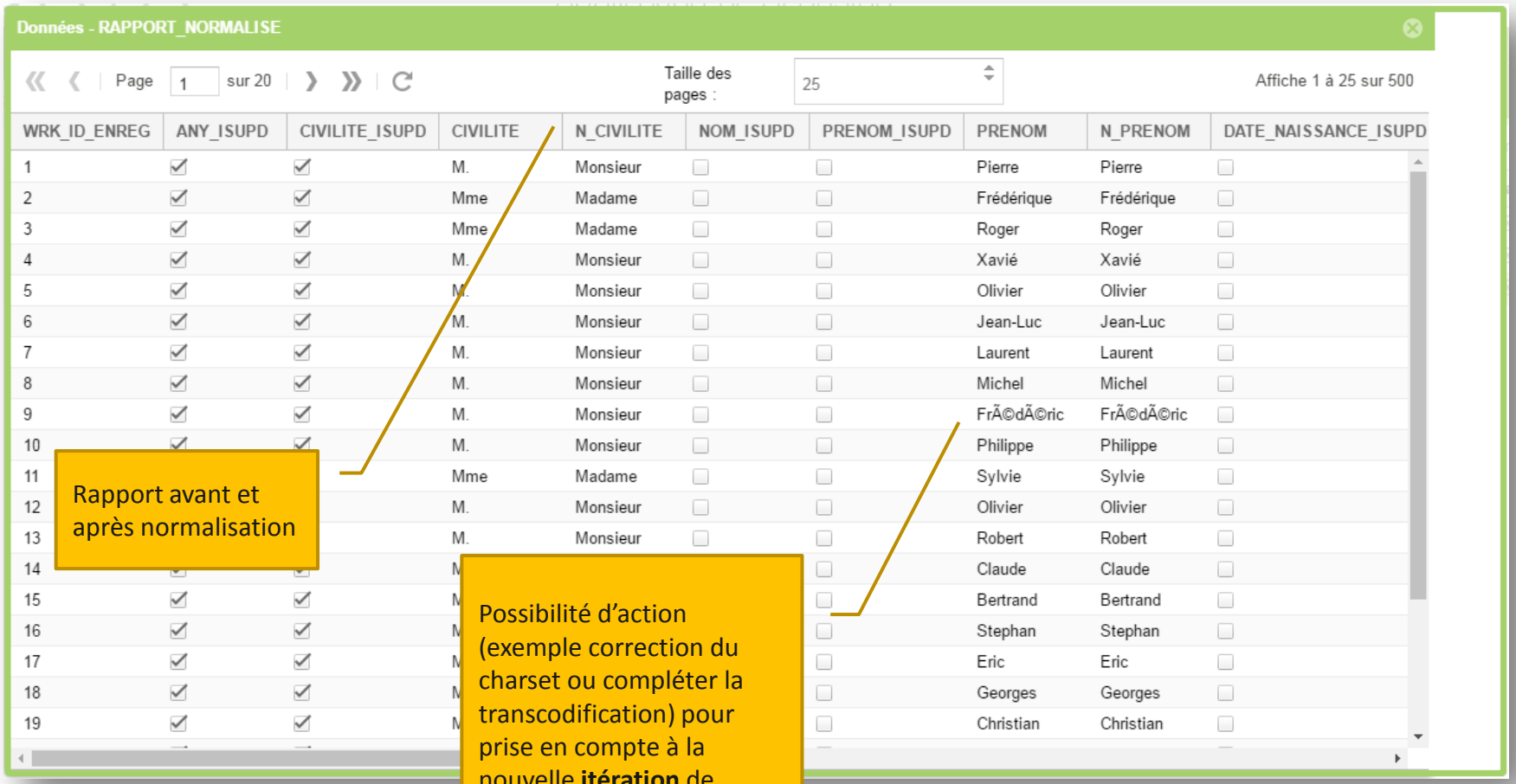


LIGNE	DOSSIER	JOB	LIBELLE_ERREUR
3	Conformité	Concordance du genre entre prénom et civilité	La valeur du champ PRENOM="Roger" et du champ CIVILITE="Madame" , ne concordent pas avec le genre
5	Conformité	Conformité du champ mail	La valeur du champ MAIL_DIRECT="olivier.bouas-laurent@@generationmedia.fr" 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 char	ECT="" n'est pas de la forme xxx@domain
309	Conformité	Conformité du char	ECT="" n'est pas de la forme xxx@domain
327	Conformité	Conformité du char	ECT="" n'est pas de la forme xxx@domain

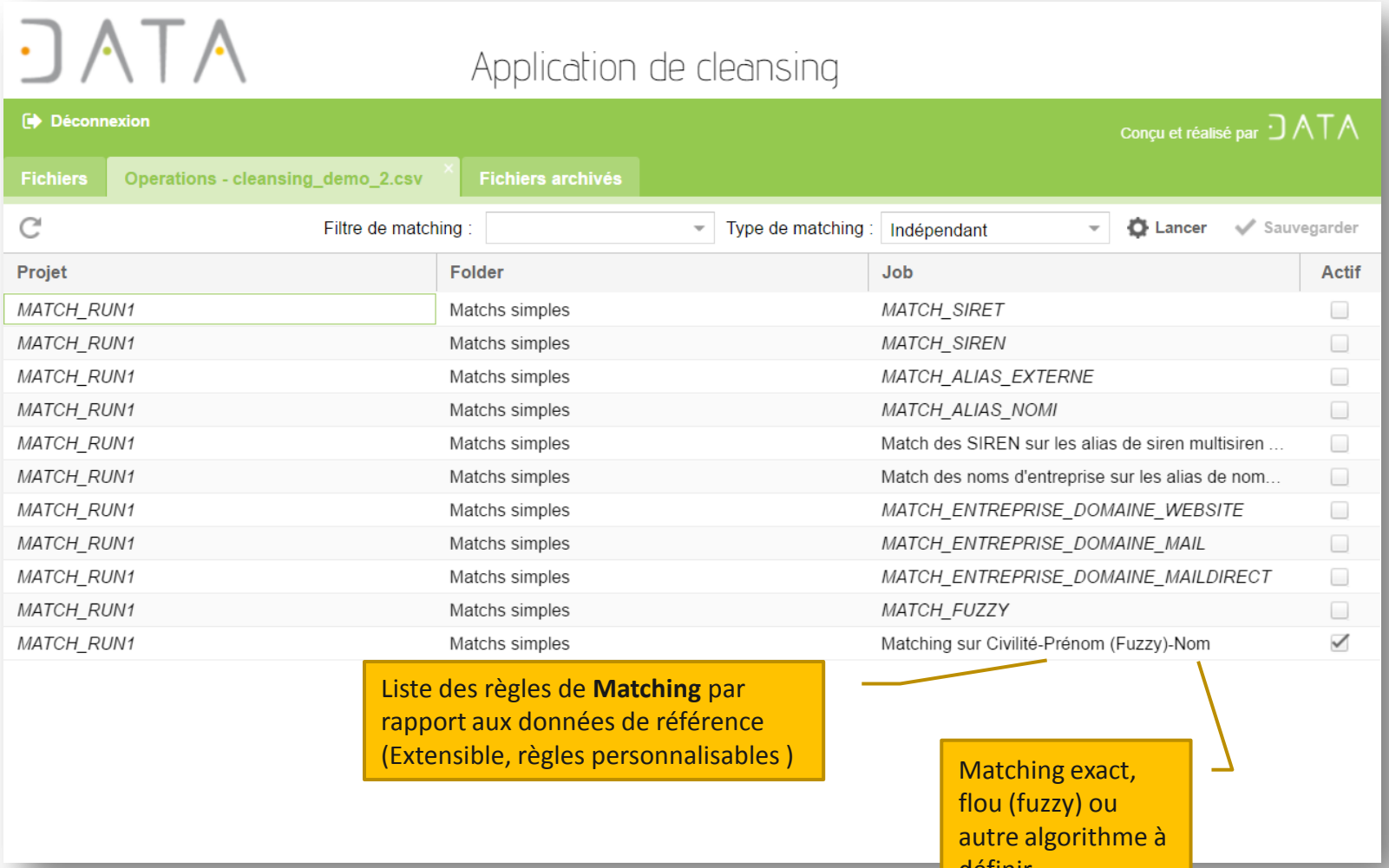
Normalisation



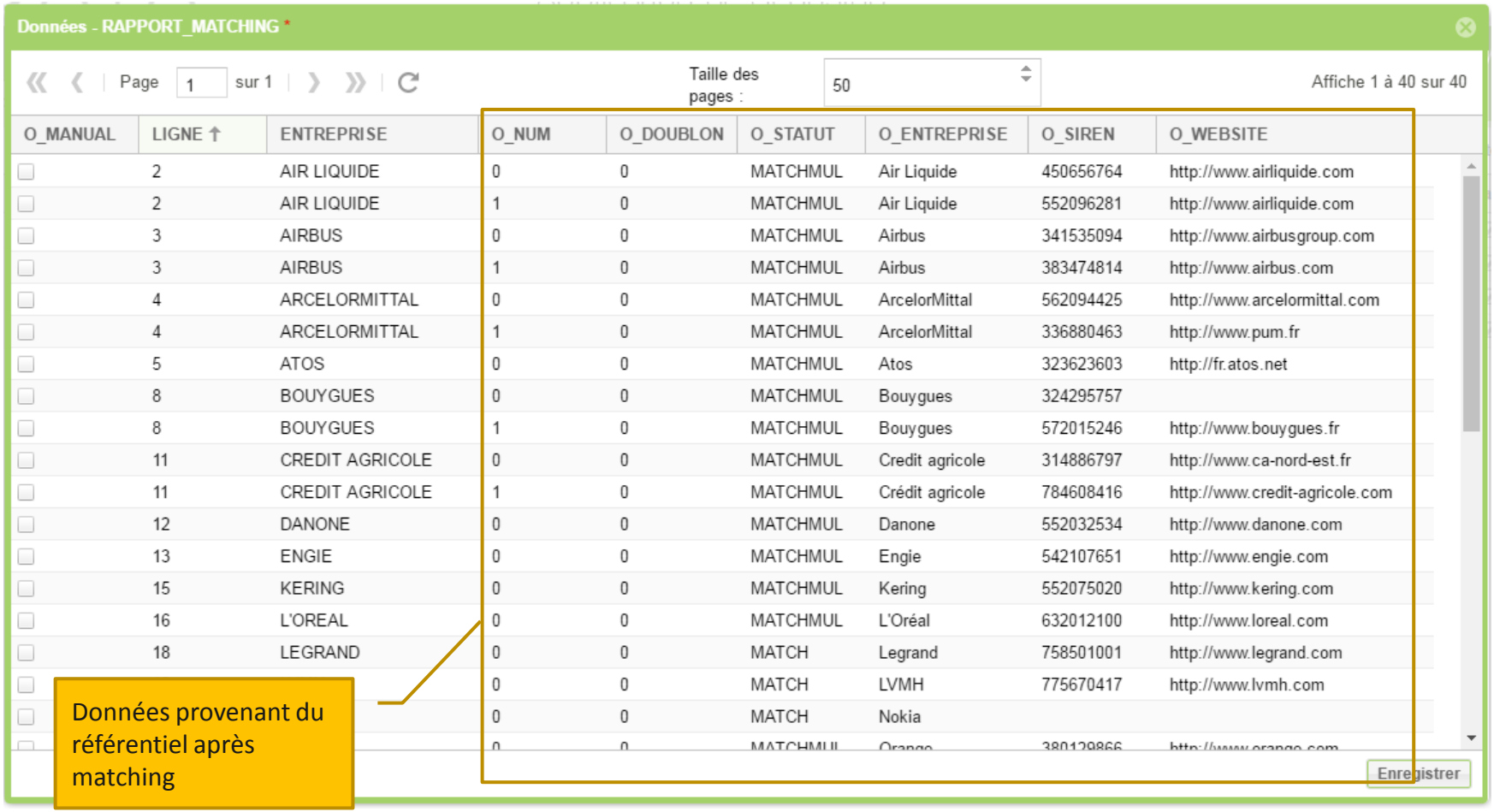
Rapport de normalisation



Matching


Application de cleansing
Déconnexion
Conçu et réalisé par
Fichiers
Operations - cleansing_demo_2.csv
Fichiers archivés
Filtre de matching :
Type de matching :
Indépendant
Lancer
Sauvegarder
Projet
Folder
Job
Actif
MATCH_RUN1 Matchs simples MATCH_SIRET
MATCH_RUN1 Matchs simples MATCH_SIREN
MATCH_RUN1 Matchs simples MATCH_ALIAS_EXTERNE
MATCH_RUN1 Matchs simples MATCH_ALIAS_NOMI
MATCH_RUN1 Matchs simples Match des SIREN sur les alias de siren multisiren ...
MATCH_RUN1 Matchs simples Match des noms d'entreprise sur les alias de nom...
MATCH_RUN1 Matchs simples MATCH_ENTREPRISE_DOMAINE_WEBSITE
MATCH_RUN1 Matchs simples MATCH_ENTREPRISE_DOMAINE_MAIL
MATCH_RUN1 Matchs simples MATCH_ENTREPRISE_DOMAINE_MAILDIRECT
MATCH_RUN1 Matchs simples MATCH_FUZZY
MATCH_RUN1 Matchs simples Matching sur Civilité-Prénom (Fuzzy)-Nom
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



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

Merci de votre attention

Retrouvez-nous sur www.data.fr