

# Outil de calcul du score composite sur la qualité des données

---

Zola Allen, Jenny Mwanza et Fredrick Onyango, Data.FI



# Aperçu

- Développement du score composite sur la qualité des données (SQD)
- Outil sur mesure de calcul du score de qualité des données

# Introduction

- Il existe un large éventail de stratégies conçues pour améliorer la qualité des données agrégées rapportées de manière systématique\*
- Nous recommandons l'utilisation d'approches mobilisant des ressources moins importantes avant de recourir à des méthodes nécessitant des ressources considérables
- Nous reconnaissons qu'une étude documentaire recourant à une notation de la qualité des données connaît certaines limites

Notation de la qualité des données

Audits de qualité des données réguliers

Audits de qualité des données

	Personnel S&E	Auditeurs externes	Visites sur le terrain	Examen documentaire	Ressources
	x			x	Faible
	x		x		Intermédiaire
		x	x		Élevé

\* Les relevés longitudinaux individuels (notamment les relevés médicaux électroniques ou RME) nécessitent des approches différentes en termes d'examen documentaire et de notation de la qualité des données (non traité ici) <sup>1</sup>

# Composantes

## COMPLÉTUDE :

mesure le nombre de dossiers soumis par rapport au nombre de dossiers attendus

TX\_CURR a été utilisé comme une mesure permettant de déterminer indirectement si une structure avait soumis les éléments de données pour une semaine précise.

TX\_CURR a été choisi dans la mesure où l'on partait du principe qu'une fois qu'une structure avait rapporté TX\_CURR, elle continuait à rapporter cet indicateur, alors que pour une semaine donnée elle ne rapportait pas forcément TX\_NEW, par exemple.

---

## COHÉRENCE :

mesure le degré selon lequel les éléments de données sont cohérents les uns par rapport aux autres

Le score de cohérence des données est utilisé pour mesurer les éventuels problèmes de qualité des données provenant du fait que le numérateur d'un indicateur est supérieur à son dénominateur. Ce score est attribué aux indicateurs suivants : HTS\_TST & HTS\_TST\_POS, TX\_NEW & TLD\_NEW, TX\_CURR & TLD\_CURR et TX\_PVLS\_D & TX\_PVLS\_N.

---

## UNIFORMITÉ :

mesure à quel point les données sont uniformes au fil du temps

Cela permet de mesurer les tendances de rendement concernant les indicateurs de rapport à haute fréquence (HFR) au cours d'une période donnée et d'identifier les valeurs aberrantes dans les rapports de données.

# Indicateurs à l'étude

En adaptant sur mesure l'approche SQD, nous collaborons avec les parties prenantes pour déterminer les indicateurs les plus significatifs à examiner selon le contexte du pays.

Dans les exemples suivants, nous avons examiné les données HFR et plusieurs combinaisons de ces indicateurs. L'outil peut être modifié pour inclure d'autres indicateurs, le cas échéant.

HTS_TST	HTS_TST_POS
TX_NEW	TLD_NEW
TX_CURR	TLD_CURR
TX_PVLS_D	TX_PVLS_N

# Scores sur la qualité des données

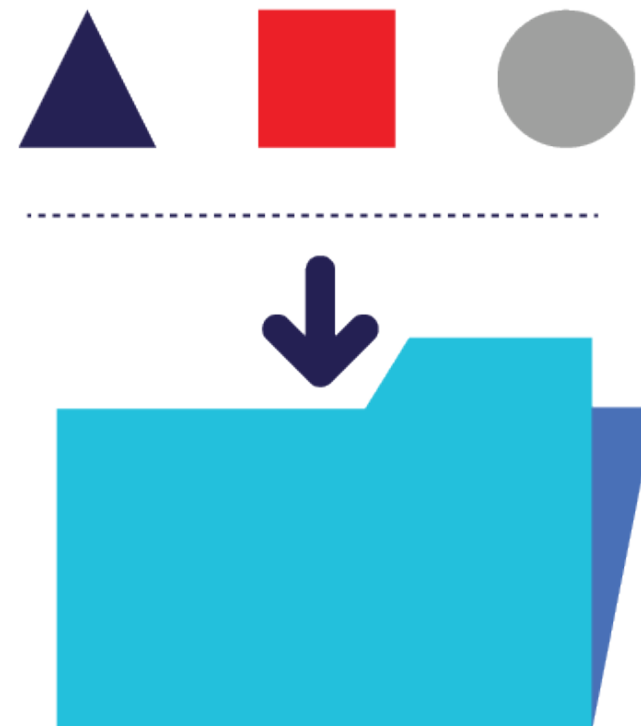
	Complétude	Cohérence	Uniformité	Score sur la qualité des données
Partenaire de mise en œuvre A	97%	0%	50%	49%
Partenaire de mise en œuvre B	93%	0%	50%	48%
Partenaire de mise en œuvre C	100%	100%	50%	83%
USAID				60%

Le score de qualité des données nous permet de quantifier le degré de performance parmi les partenaires de mise en œuvre, les unités organisationnelles infranationales ou les pays au sein d'une région.

Le principe directeur est que lorsque les scores sont transparents et calculés de manière uniforme, on observe un niveau de redevabilité accru. De plus, les parties responsables chercheront à améliorer les scores.

# Complétude

---



# Méthodes

- HTS\_TST a été utilisé comme une mesure permettant de déterminer indirectement si une structure avait soumis les éléments de données pour une semaine précise.
- Rapport attendu = Nombre total de structures x nombre de rapports pour la période.
- Les rapports reçus représentent la somme de tous les rapports pour la période (semaines).
- Le score en pourcentage est déterminé en divisant le nombre de rapports reçus par le nombre de rapports attendus x 100.

$$\frac{\text{Nombre de rapports reçus}}{\text{Nombre de rapports attendus}} \times 100$$



# Scores de complétude

Partenaire de mise en œuvre	Structures de santé	Nombre de semaines	Nombre de rapports attendus	Nombre de rapports reçus	Score
Partenaire A	25	16	400	388	97%
Partenaire B	17	16	272	202	74%

# Cohérence

---



# Méthodes

**COHÉRENCE :** mesure le degré selon lequel les éléments de données sont cohérents les uns par rapport aux autres selon les mesures des règles de validation.

- HTS\_TST\_POS  $\leq$  HTST\_TST
  - TLD\_NEW  $\leq$  TX\_NEW
  - TLD\_CURR  $\leq$  TX\_CURR
  - TX\_PVLS\_N  $\leq$  TX\_PVLS\_D
- Chaque paire de règles de validation reçoit un score maximum de cohérence en pourcentage qui est calculé comme 100% divisé par le nombre de règles.
    - S'il y a quatre règles, il est attribué à chaque règle de validation un score de cohérence de 25%.
  - Ce score maximum est conservé si une règle de validation est satisfaite par toutes les structures. Un score de zéro est attribué si au moins une des structures n'a pas satisfait à une règle de validation.
  - Le score total de cohérence est calculé comme une somme de tous les scores des règles de validation.
  - Chaque paire de règles de validation ayant été validée doit nécessairement être corrigée.

# Scores de cohérence

Cohérence	$\text{HTS\_TST} \geq \text{HTS\_TST\_POS}$	$\text{TX\_NEW} \geq \text{TLD\_NEW}$	$\text{TX\_CURR} \geq \text{TLD\_CURR}$	$\text{TX\_PVLS\_D} \geq \text{TX\_PVLS\_N}$	Score
Partenaire de mise en œuvre A	3	52	5	17	0%
Partenaire de mise en œuvre B	4	13	13	1	0%
Partenaire de mise en œuvre C	0	0	0	0	100%

- Le score de cohérence des données est utilisé pour mesurer les éventuels problèmes de qualité des données provenant du fait que le numérateur d'un indicateur est supérieur à son dénominateur.
- Il est attribué à chaque règle de validation un score de cohérence de 25%

# Exemples de cohérence

DÉPISTAGE		
TX_PVLS_D	TX_PVLS_N	TX_PVLS_D ≥ TX_PVLS_N
135	149	<b>VIOLATION</b> Le nombre de patients avec un test mesurant la suppression de la charge virale doit être égal ou supérieur au nombre de patients dont la charge virale a été supprimée.

TRAITEMENT		
TX_NEW	TLD_NEW	TX_NEW ≥ TLD_NEW
2	6	<b>VIOLATION</b> Le nombre de patients ayant initié un traitement TAR doit être égal ou supérieur au nombre de patients ayant initié le régime thérapeutique TLD.

Les tableaux ci-dessus avec les cellules surlignées montrent les indicateurs avec des problèmes de qualité des données en termes de cohérence.

# Scores de cohérence | Étude de base

ÉTUDE DE BASE											
Pays	HTS_TST_POS > HTS_TST					Indice HTS_TST_POS > Indice HTS_TST					Score
	F <15	H <15	F 15+	H 15+	TOUS	F <15	H <15	F 15+	H 15+	TOUS	
TOGO	0	0	0	0	0	0	1	1	2	0	70%
BURKINA	0	0	0	0	0	0	1	2	1	1	60%

**Il est attribué à chaque règle de validation un score de cohérence de 10% (il y a 10 règles de validation dans cet exemple).**

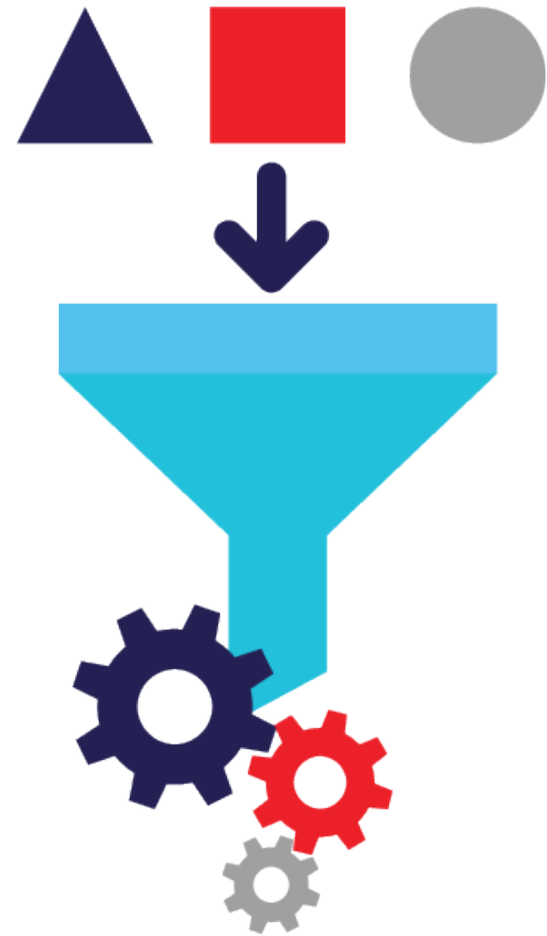
# Structures avec des problèmes de cohérence, Togo | Étude de base

CMS Lucia (EVT)   Étude de base										
Date	HTS_INDEX_TST					HTS_INDEX_POS				
	F < 15	M < 15	F 15+	H 15+	TOUS	F < 15	M < 15	F 15+	H 15+	TOUS
13/04/2020	2	0	2	0	4	0	1	0	1	2

Hôpital de Bè   Étude de base										
Date	HTS_INDEX_TST					HTS_INDEX_POS				
	F < 15	M < 15	F 15+	H 15+	TOUS	F < 15	M < 15	F 15+	H 15+	TOUS
29/06/2020	4	0	0	3	7	0	0	3	3	6
18/05/2020	0	4	3	0	7	0	0	1	1	2

# Uniformité

---





# Méthodes

Il s'agit d'une mesure d'**UNIFORMITÉ** dans les relevés au fil du temps ou simplement de l'absence de valeurs aberrantes :

1. TX\_NEW
  2. TX\_CURR
- Les valeurs identifiées ne correspondent pas à la tendance rapportée par cette structure.

# Méthodes (suite)

- Identifiez la présence de valeurs aberrantes :
  - Pour chaque indicateur, calculez les valeurs minimales et maximales au cours d'une période de temps donnée.
  - Utilisez le formatage conditionnel pour souligner qu'une structure affiche des valeurs aberrantes si la valeur maximale est au moins cinq fois supérieure à la valeur minimale.
  - Faites en sorte qu'un individu examine toutes les valeurs soulignées et déterminez au final quelle structure présente des valeurs aberrantes sur la base des directives fournies.
- Chaque indicateur mesurant le degré d'uniformité reçoit un score maximum qui est calculé comme 100% divisé par le nombre d'indicateurs examinés.
  - S'il y a deux indicateurs, il est attribué à chaque mesure un score d'uniformité de 50%.

# Méthodes (suite)

- Le score maximum est conservé s'il n'y a pas de valeurs aberrantes dans une quelconque structure. Un score de zéro est attribué si au moins une des structures affiche des valeurs aberrantes.
- Le score total d'uniformité est calculé comme la somme des scores d'uniformité de tous les indicateurs.

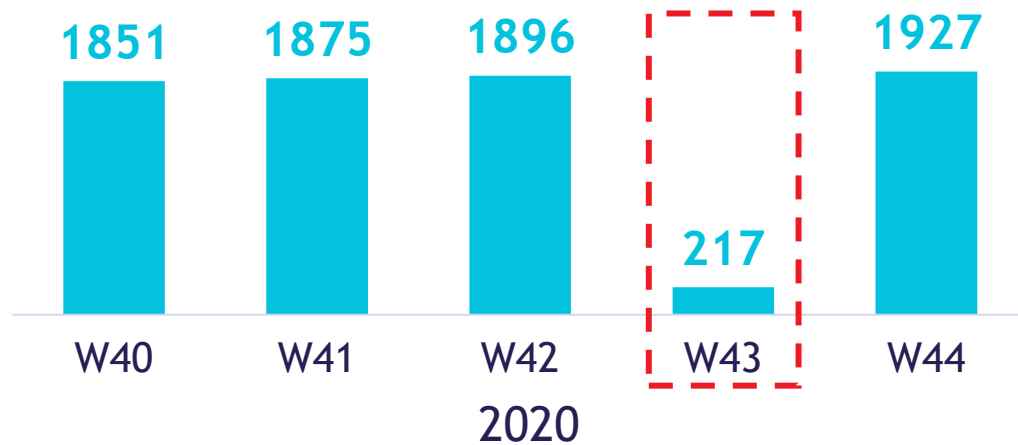
# Scores d'uniformité

Uniformité	TX_NEW	TX_CURR	Score
Partenaire de mise en œuvre A	2	0	50%
Partenaire de mise en œuvre B	0	5	50%
Partenaire de mise en œuvre C	0	2	50%

Chaque mesure d'uniformité sur un indicateur spécifique reçoit un score de 50%.

# Exemples d'uniformité

Hôpital général A

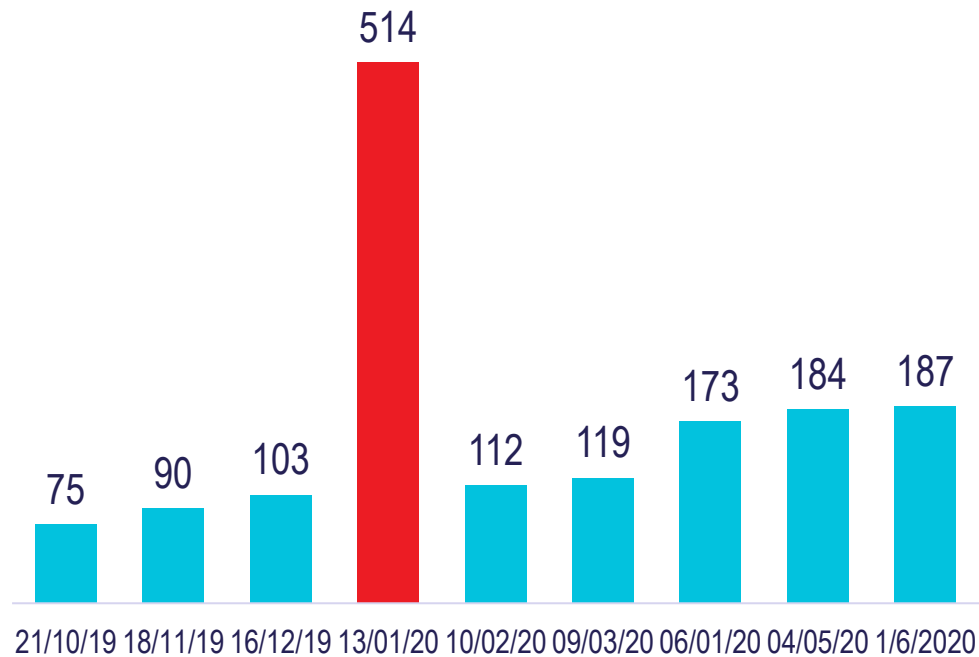


Hôpital général B

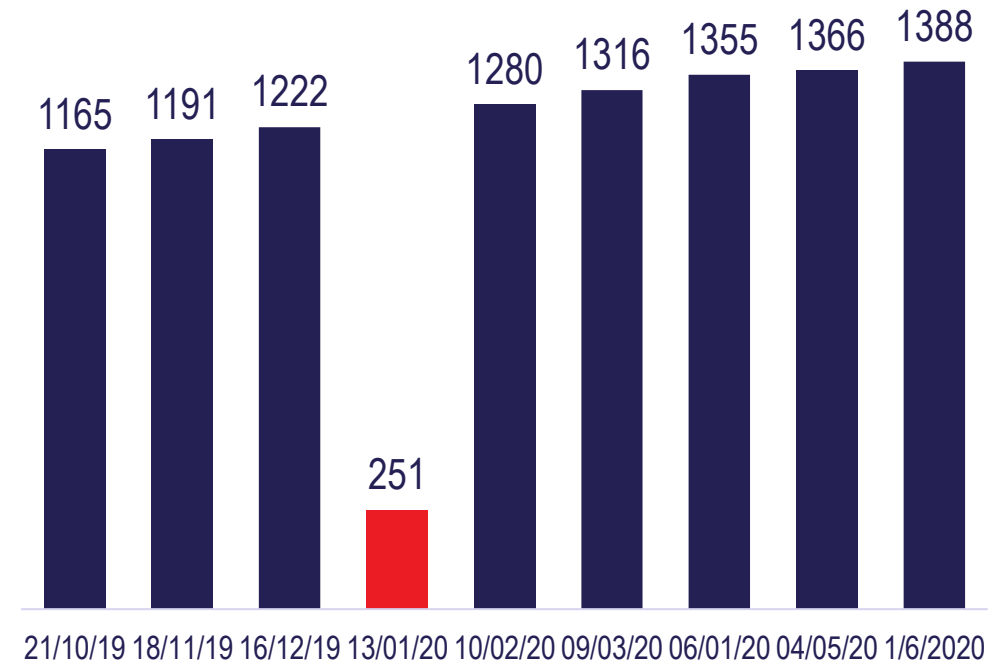


# Structures avec des problèmes d'uniformité, TX\_CURR | Étude de base

Hôpital A

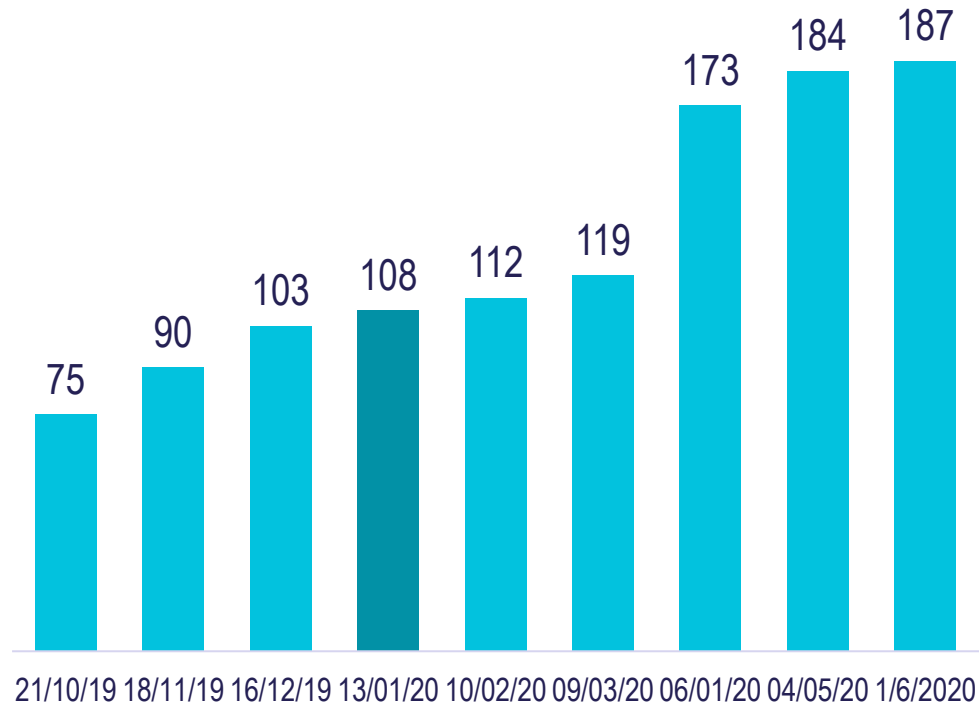


Hôpital B

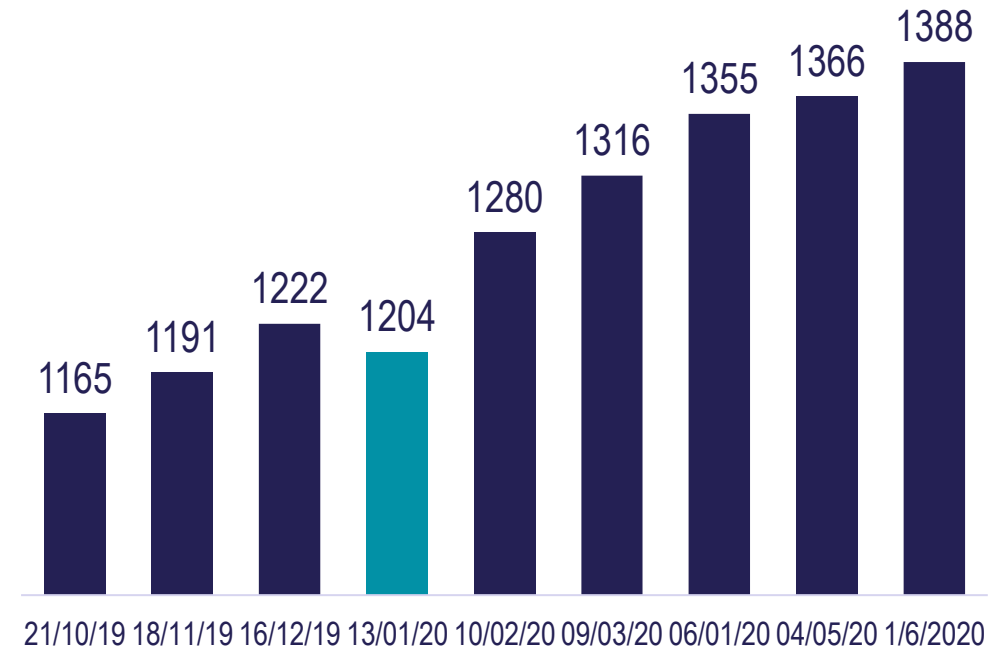


# Structures avec des problèmes d'uniformité, TX\_CURR | Étude finale

Hôpital A

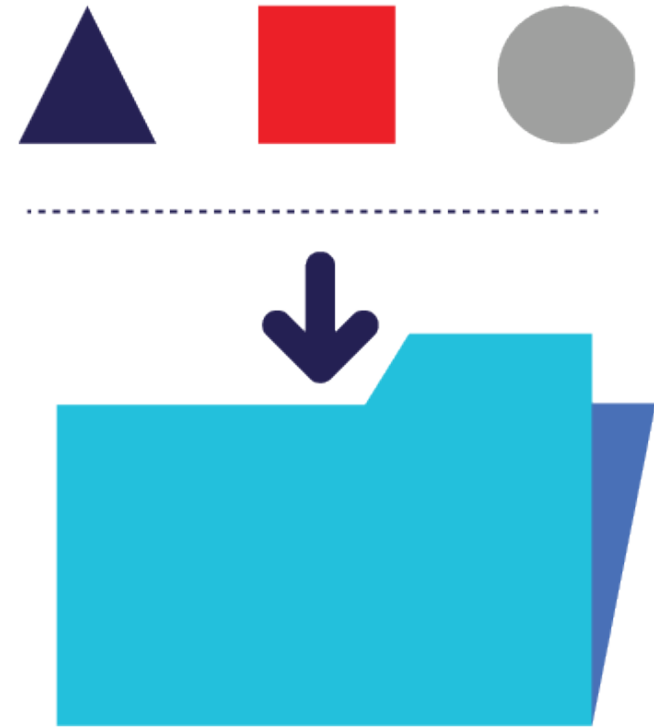


Hôpital B



Outil adapté sur  
mesure pour calculer  
le score de qualité  
des données

---





# Sélection de la plateforme

Initialement, nous avons calculé manuellement le score de qualité des données, soit un processus fastidieux et laborieux. Au moment de choisir un outil pour automatiser ce processus, notre équipe a examiné les outils existants suivants :

1. Data Cleaner - <http://datacleaner.org/>
2. DataPreparator - <http://www.datapreparator.com/>
3. Data Quality Analyzer - <https://www.ataccama.com/product/data-discovery-and-profiling/dqa>
4. WHO DHIS2 Data Quality Tool - <https://who.dhis2.org/dq/dhis-web-commons/security/login.action>

# Critères de sélection

Exigences concernant l'outil
Extrait et partage des résumés des scores de qualité des données
Extrait et partage des données brutes évaluées pour déterminer la qualité des données
Capacité à importer automatiquement de nouvelles données (Excel) fournies et à calculer automatiquement des scores sans apport manuel supplémentaire
Mesure de la complétude – Degré selon lequel toutes les données requises pour un échantillon donné sont disponibles dans l'ensemble de données ; une mesure du pourcentage des saisies de données depuis les unités
Mesure de la cohérence – Degré selon lequel les données adhèrent à des règles d'entreprise
Mesure de l'uniformité – Mesure des valeurs aberrantes
Minimise le risque de doublons et d'erreurs, tout en maintenant l'intégrité des données
L'outil proposé doit être solide et être doté de critères de notation adaptables pour mesurer des aspects différents/supplémentaires à long terme
Capacité à stocker des données au moins jusqu'à la fin de sa période d'utilité
Capacité à importer et exporter des données, au moins sous Excel – capacité à partager des rapports le cas échéant
Simple à utiliser
Facile à installer
Exclusion des valeurs nulles

# Développement d'un outil sur mesure

Lorsqu'aucune des plateformes de qualité des données existantes ne répond aux critères de sélection, Microsoft Excel avec l'application Visual Basic (VBA) a été retenu.

Pour utiliser l'Outil de notation de la qualité des données découlant de cette démarche, suivez les étapes suivantes :

1. Préparez des fichiers de données brutes
2. Importez des fichiers de données brutes
3. Exécutez le script
4. Examinez les résultats



L'outil produit aussi bien des rapports détaillés que des analyses succinctes qui peuvent être facilement partagées avec les parties prenantes

L'outil peut contenir jusqu'à 10 000 relevés

- Jusqu'à 12 périodes de rapport
- Jusqu'à 10 indicateurs
- Jusqu'à 80 unités de rapport



L'outil peut facilement être adapté pour des ensembles de données autres que les données HFR du PEPFAR. Le nombre d'applications possibles est illimité !

# Application SQD actuelles

- SQD calculé en utilisant les données des rapports à haute fréquence sur le VIH
  - Niger, Mali, Ghana, Togo, Burkina Faso, Liberia, Sénégal
- Indicateurs inclus
  - HTS\_TST, HTS\_TST\_POS
  - HTS\_Index\_TST, HTS\_Index\_TST\_POS
  - KP\_PREV
  - TX\_NEW
  - TX\_CURR
- Score de base moyen à travers les pays – 74%
- Score final moyen – 100%

# Application SQD actuelles (suite)

Pays	Période examinée	Nombre de semaines à examiner	Nombre de structures de santé incluses	Domaines nécessitant une validation ou un nettoyage des données		
				Complétude	Cohérence	Uniformité
Ghana	30 septembre-6 juin 2020	37	25		X	X
Burkina Faso	16 mars-29 juin 2020	16	17		X	X
Liberia	20 janvier-29 juin 2020	24	18		X	
Mali	16 décembre-2 mars 2020	12	13		X	
Sénégal	20 janvier-1 <sup>er</sup> juin 2020	12	8		X	
Togo	16 mars-29 juin 2020	16	25		X	X
Nigeria	30 décembre-23 mars 2020	13	479	X	X	X

# Résultats du processus de sélection (Annexe)

Requirements	<a href="#">Data Cleaner</a>	<a href="#">DataPreparator</a>	<a href="#">Data Quality Analyzer</a>	<a href="#">WHO DHIS Quality Tool</a>
Extract and share summary data quality scores	No	Install failed	Yes	Yes
Extract and share raw data assessed for data quality	Yes		Yes	Yes
Ability to automatically import new (excel) data provided and auto compute scores without further manual input.	Yes		No	Yes - Must be prepared into DHIS2 import format
Measure Completeness : degree to all required data are available in the dataset. A measure of the percentage of expected data entries from units	No		No	No
Measure Coherence : Extent to which data adheres to the business rules	Yes		Yes	No
Measure Consistency : Measure outliers. Data should be constant in time	No		No	Yes
Minimize risk of duplication, errors and maintain integrity	Yes		Yes	Yes
Proposed tool should be robust with customizable scoring criteria to cater for different/additional measurements in the long run	Yes - Not fully customizable		Yes - To some extent	No
Ability to store data at least until utility is complete.	Yes		Yes	Yes
Data import and export at least in excel. Ability to share report as required	Yes		No	Yes
User Friendly	To some extent		No	Yes
Easy to Install	Yes	No	Yes	No
Exclude Null Values	Yes		Yes	Yes



## POUR EN SAVOIR PLUS

Emily Harris, AOR Data.FI, Bureau de  
l'USAID pour la lutte contre le VIH/SIDA  
[emharris@usaid.gov](mailto:emharris@usaid.gov)

Jenifer Chapman, Directrice du projet Data.FI  
[datafiproject@thepalladiumgroup.com](mailto:datafiproject@thepalladiumgroup.com)

Le projet Data for Implementation (Data.FI) est un accord de coopération quinquennal financé par le Plan d'urgence du Président des États-Unis en matière de lutte contre le SIDA, par le biais de l'Agence des États-Unis pour le développement international, aux termes de l'accord N°7200AA19CA0004 entré en vigueur le 15 avril 2019. Il est mis en œuvre par Palladium, en partenariat avec le JSI Research & Training Institute (JSI), le Département d'épidémiologie de Johns Hopkins University (JHU), Right to Care (RTC), Cooper/Smith, IMC Worldwide, Jembi Health Systems et Macro-Eyes, et appuyé par des partenaires-ressources experts au niveau local.

Cette présentation a été produite en vue d'un examen par le Plan d'urgence du Président des États-Unis en matière de lutte contre le SIDA, par le biais de l'Agence des États-Unis pour le développement international. Elle a été préparée par le projet Data for Implementation. Les renseignements fournis dans cette présentation ne sont pas des informations officielles émanant du gouvernement américain et ne représentent pas nécessairement les opinions ou positions du Plan d'urgence du Président des États-Unis en matière de lutte contre le SIDA, de l'Agence des États-Unis pour le développement international ou du gouvernement américain.