

Les plateformes des données ouvertes

15 Décembre 2021 – Séance n°5



UNITED NATIONS

الاسواق
ESCWA

Shared Prosperity **Dignified Life**



La qualité des données

Une grande qualité des données ?

- Les données sont jugées de grande qualité si elles représentent correctement la réalité à laquelle elles se réfèrent.
- Le degré de pertinence / conformité des données aux usages prévus.

La qualité des données peut être dégradée à deux niveaux :

- Au niveau des métadonnées (descriptives, etc...)
- Au niveau est celui de la donnée en elle-même

Évaluation de la qualité des données :

- Évaluation subjective (subjective = dépend de la personne)
- Évaluation objective (conformité par rapport aux normes aux référentiels)

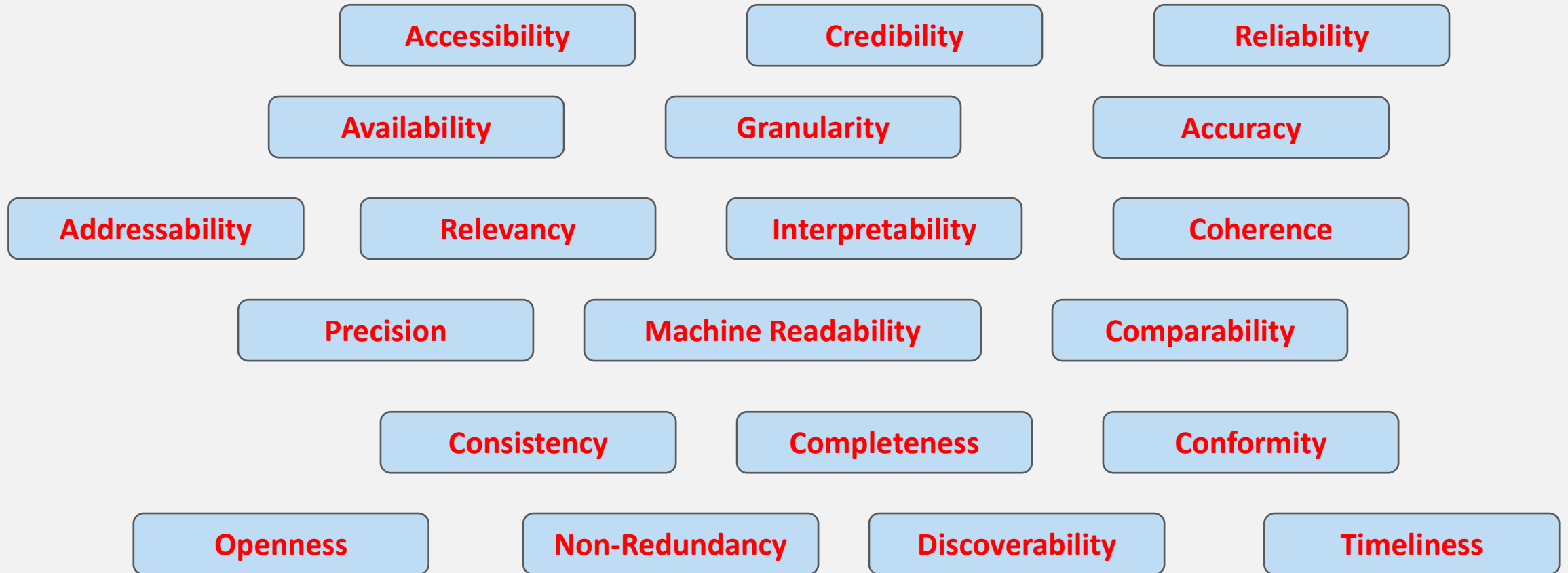
La qualité des données ouvertes

```
:stops-2015-05-05.csv
  a dcat:Distribution ;
  dcat:downloadURL <http://data.mycity.example.com/transport/dataset/bus/stops-2015-0
dct:title "CSV distribution of stops-2015-05-05 dataset" ;
dct:description "CSV distribution of the bus stops dataset of MyCity" ;
dcat:mediaType "text/csv;charset=UTF-8" ;
dct:license <http://creativecommons.org/licenses/by-sa/3.0/> ;
dqv:hasQualityMeasurement :measure1, :measure2
.
:measure1
  a dqv:QualityMeasurement ;
  dqv:computedOn :stops-2015-05-05.csv ;
  dqv:isMeasurementOf :downloadURLAvailabilityMetric ;
  dqv:value "true"^^xsd:boolean
.
:measure2
  a dqv:QualityMeasurement ;
  dqv:computedOn :stops-2015-05-05.csv ;
  dqv:isMeasurementOf :csvCompletenessMetric ;
  dqv:value "0.5"^^xsd:double
.
#definition of dimensions and metrics
:availability
  a dqv:Dimension ;
  skos:prefLabel "Availability"@en ;
  skos:definition "Availability of a dataset is the extent to which data (or some por
  dqv:inCategory :accessibility
.
:completeness
  a dqv:Dimension ;
  skos:prefLabel "Completeness"@en ;
  skos:definition "Completeness refers to the degree to which all required informati
  dqv:inCategory :intrinsicDimensions
.
:downloadURLAvailabilityMetric
  a dqv:Metric ;
  skos:definition "It checks if dcat:downloadURL is available and if its value is der
  dqv:inDimension :availability
.
:csvCompletenessMetric
  a dqv:Metric ;
  skos:definition "Ratio between the number of objects represented in the cvs and the
  dqv:inDimension :completeness
.
.
```

Data Quality - CSV Distribution

Completeness	Refers to the degree to which all required information is present in a particular dataset.
Metric	Ratio between the number of objects represented in the cvs and the number of objects expected to be represented according to the declared dataset scope.
Value	0.5 (Double)
Availability	Refers to the extent to which data (or some portion of it) is present, obtainable and ready for use.
Metric	dcat:downloadURL is available and its value is dereferenceable.
Value	true(Boolean)

La qualité des données ouvertes



La qualité des données ouvertes



La qualité des données ouvertes

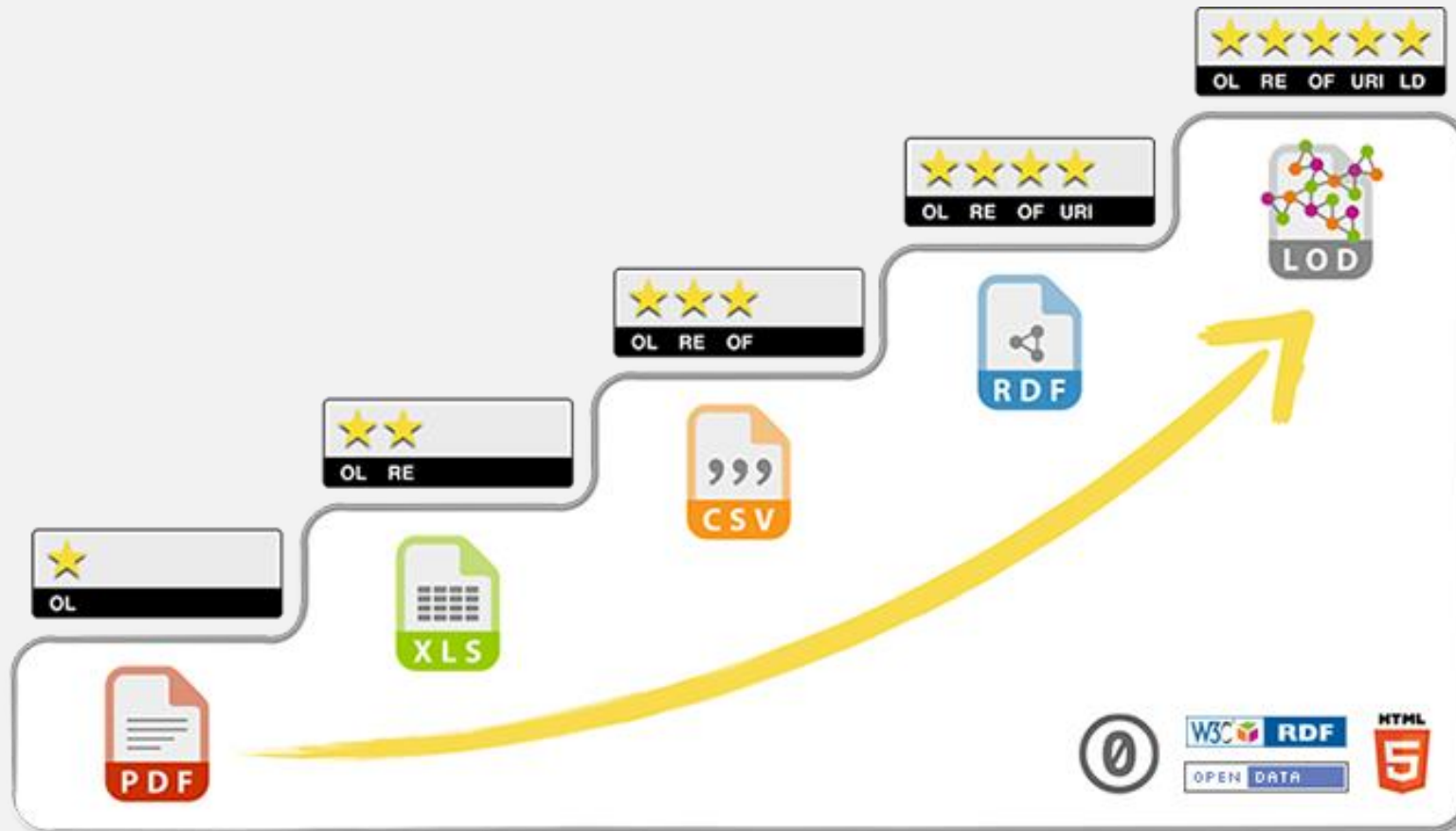
Règle 80:20 lors de la réutilisation des données ouvertes ?

80 % du temps pour améliorer la qualité des données et seulement 20 % pour analyser et comprendre les données en vue de leur réutilisation ?

Les mécanismes de l'évaluation de la qualité des données ouvertes ?

- Modèle de 5 étoiles de TBL
- Certificat de données ouvertes "Open Data Certificate"
- Modèle Français de Qualité Open Data (Sprint Qualité)
- Cadre Procédural de l'évaluation et de l'amélioration de la qualité des données ouvertes en Jordanie

Modèle de 5 étoiles de TBL



Open Data Certificate

Les 'Open Data Certificates' de l'Open Data Institute sont des exigences pratiques, sociales, légales et techniques à prendre en compte lors de la publication ou de l'utilisation d'open data.

<https://certificates.theodi.org/en/>

- **BRONZE** Les données publiées sont sous licence ouverte permettant leur réutilisation
- **ARGENT** : BRONZE + données pouvant être traitées automatiquement en plus de l'existence de la documentation nécessaire
- **GOLD** : ARGENT + conformité et utilisation des standards, tout en assurant des mises à jour périodiques
- **PLATINUM** : GOLD + Ajout d'URI et existence d'une équipe pour répondre aux questions sur les données ouvertes

Sprint Qualité

Sprint Qualité ?

Un exercice qui dure entre deux heures et demie et trois heures pour un jeu de données ouvertes ?

Au moins Trois personnes ?

120 questions pour vérifier la qualité des données ?

Une version allégée basée sur seulement 10 questions !

Sprint Qualité

I D	Type de problème	Observateurs	Problème rencontré ?	Commentaire + exemple réel dans le jeu de données	A quel niveau ou fréquence le problème est-il observé ?	Impact du problème sur les usages	Effort de correction estimé
39	Manque de métadonnées : la fraîcheur des données n'est pas explicité : * le délai entre le réel et la mise en base de la donnée * le délai entre le réel et la publication de la donnée	Binôme 4	oui	on a effectivement une date de mise à jour cette date n'est pas necessairement identique à la date réel de modification de l'équipent	1. concerne l'ensemble du jeu	important	important
40	Manque de métadonnées : la langue des textes n'est pas spécifiée	Binôme 4	oui	suffisamment parlant	1. concerne l'ensemble du jeu	nul ou négligeable	très faible
41	Métadonnées imprécises : le format de date n'est pas spécifié	Binôme 4	oui	pas de mention du format	1. concerne l'ensemble du jeu	faible	faible
42	Métadonnées imprécises : unités non spécifiées	Binôme 4	non	unité précisée en M2			
43	Métadonnées imprécises : système de coordonnées non spécifié	Binôme 5	non	pas de coordonnée GPS			
44	Métadonnées imprécises : nom de colonnes ambigus	Binôme 5	non	nom des colonnes décrit dans méta données			
45	Métadonnées fausses	Binôme 5	oui	incomplètes : exemple manque Insnovoie	4. problème fréquent	faible	très faible
46	La taille maximale d'un champ dépasse celle qui est spécifiée dans la documentation	Binôme 5	oui	taille non indiquée dans les méta données : ne permet pas de contrôler	4. problème fréquent	faible	
47	L'ordre des colonnes ne correspond pas à l'ordre donné dans la documentation	Binôme 5	oui	on a été obligé de regrouper et trier les méta données de 3 onglets	4. problème fréquent	faible	très faible

OGDQ Jordanie

Les dimensions d'évaluation et de mesure de la qualité des données ouvertes :

Phase de Prépublication

Étape proactive

- Conformité légale

Étape de priorisation

- Importance des données

Phase de la 1^{ère} publication

- Validité des données
- Complétude des données
- Accessibilité aux données
- La granularité des données
- Des temps à temps
- Ouverture des données
- Exploitation automatique des données

Phase de la mise à jour

- Validité des données
- Complétude des données
- Accessibilité aux données
- La granularité des données
- Des temps à temps
- Ouverture des données
- Exploitation automatique des données
- Feedback

OGDQ Jordanie

Les indicateurs d'évaluation et de mesure de la qualité des données ouvertes :

Phase de Prépublication

Étape proactive

Étape de priorisation

4 indicateurs

5 indicateurs

Phase de la 1^{ère} publication

21 indicateurs

Phase de la mise à jour

24 indicateurs



Shared Prosperity **Dignified Life**



Discussion