

Les dimensions Qualité

Etat de l'Art

Groupe de travail Exqivalence

Avant propos

Qu'est ce que la qualité?

Définition de l'ISO (1986)

“the totality of features and characteristics of an entity that bears on its ability to satisfy stated and implied needs”

⇒ l'ensemble des caractéristiques requises d'une entité qui font qu'elle soit conforme aux différents besoins

Qu'est-ce qu'une dimension qualité?

Une dimension qualité est une caractéristique qui définit une propriété des données [Batini 06].

La sélection des dimensions qualité et leurs définitions diffèrent d'une problématique à une autre et d'un contexte à un autre

On parle aussi de critères qualité, facteurs qualités, ...

La qualité partout...

Les dimensions qualité ont été définies dans plusieurs contextes

- les données géographiques
- le domaine du marketing (données statistiques)
- les télécommunications
- le domaine médical
- etc.

Données géographiques

La généalogie (ou la provenance): critère qualitatif retraçant la vie du jeu de données, depuis sa création jusqu'à la mise à disposition de l'utilisateur.

La précision:

- **Précision géométrique:** exactitude spatiale
 - Précision de position: l'objet est plus ou moins positionné sur la carte
 - Précision de forme: la forme de l'objet est plus ou moins juste sur la carte
- **Précision sémantique:** est la différence entre la valeur d'un attribut du jeu de données et sa valeur dans le monde nominal

L'exhaustivité: indique si les objets du terrain nominal sont tous représentés dans le jeu de données

La cohérence logique: le degré de cohérence interne des données selon les règles de spécifications et de modélisation du jeu de données

La cohérence sémantique: la qualité avec laquelle les objets géographiques sont décrits

La promptitude: renseigne sur la fraîcheur des données

La fidélité textuelle: l'exactitude de l'orthographe des informations écrites

Données marketing

Complétude: données manquantes dans une table/colonne, par rapport à une population de référence

Exactitude: distance entre la valeur v et la valeur v' considérée comme la représentation exacte de la réalité dont v est le représentant

Cohérence: respectant les contraintes d'intégrité

Pertinence: degré d'utilité de la donnée et son adéquation aux besoins des utilisateurs

Conformité: conformité des valeurs des données au format spécifié

Duplication: existe-t-il des doublons dans la base (représentations inutiles d'une même entité dans l'ensemble des données) => signe de vulnérabilité du système

Intégrité: les relations sont-elles toutes représentées dans le système

Utilisabilité (actionability/marketability): par exemple volume d'adresses inutilisables

Données télécommunications

Accessibilité: de point de vue utilisateur

- Disponibilité
- Sécurité

Interprétabilité: langage de description des données prenant en compte les aspects spécifiques au domaine

- Syntaxique
- Sémantique

Contexte

- Volume des données
- Pertinence
- Actualité: non volatilité VS récence

Fiabilité

- Complétude
- Exactitude
- Cohérence
- Crédibilité

On trouve aussi

- Prix des données, traçabilité/Auditabilité

Données médicales

Pour certains

- **Exactitude**: fiabilité, couverture, collection et stockage des données, estimation, imputation des données
- **Promptitude**
- **Comparabilité**: compréhensibilité, intégration, standardisation, équivalence, capacité de rapprochement, comparabilité des données produites et de l'historique
- **Utilisabilité**: accessibilité, documentation, interprétabilité
- **Pertinence**: adaptabilité, valeur

Pour d'autres

- **Intégrité**: permettant l'identification administrative des patients en toute confiance (et non précision ou exactitude): répond à la question: « quand, comment, qui et par qui ces données ont pu être générées et maintenues? »
- **Confidentialité**

Conclusion

- Plusieurs dimensions
 - 179 dimensions ont été définies en 1996 [\[Wang et Strong 96\]](#)
 - différentes d'un domaine à l'autre
 - dont la définition change d'un domaine à l'autre
- Quelle terminologie utiliser?
- Quelles dimensions garder/regrouper?
- Dans la suite, on définira les dimensions les plus importantes et les plus communément employées.

Les dimensions les plus répandues



Sommaire

Exactitude (accuracy)

Complétude (completeness)

Les dimensions temporelles:

- Promptitude (timeliness)
- Actualité (currency)
- Récence (recency)
- Volatilité (volatility)
- Fraîcheur (freshness)
- Obsolescence (obsolescence)

Cohérence (Consistency)

Les dimensions de confiance

- Crédibilité (believability)
- Objectivité (objectivity)

Précision (precision)

Pertinence (relevancy/relevance)

Unicité (uniqueness)

Exactitude (Accuracy) (1)

Porte sur la conformité de l'information contenue dans la base de données à la réalité qu'elle doit mesurer.

[Batini et Scannapieco 06] : c'est la proximité d'une valeur v à une valeur v' , considérée comme la représentation correcte d'une entité réelle que v tente de représenter

– On distingue

- L'**exactitude syntaxique**: l'admissibilité de la valeur dans le domaine de définition de l'attribut.
 - Exemple: v =John (valeur réelle), si v' =Jack, v' est considérée correcte, car admissible dans le domaine de définition (nom des personnes)
- L'**exactitude sémantique**: déviation par rapport à la valeur réelle
 - Exemple: Dans le cas ci-dessus, v' est considéré incorrect

[Naumann et al. 99]: c'est le taux des objets correctement représentés (ne comportant pas de fautes d'orthographe, valeurs appartenant au domaine de définition, ...)

Exactitude (Accuracy) (2)

[Redman 96]: “The degree of agreement between a collection of data values and a source agreed to be correct”

→ le taux de conformité d'un ensemble de valeurs de données à un ensemble de données de référence (conventionnellement correct)

[Peralta 06]: le terme exactitude est souvent utilisé pour décrire plusieurs aspects liés à la donnée intrinsèque: exactitude sémantique, syntaxique, précision, absence d'ambiguïté...

[Wang et Strong 96]: une donnée exacte est correcte, objective et extraite d'une source fiable

→ Les définitions sont différentes selon le système dans lequel évoluent les données

→ De plus, comme l'exactitude est fortement dépendante de quelques autres dimensions qualité (tq: complétude, cohérence, fraîcheur).

→ L'exactitude peut être facilement mesurée dans certains cas (présence de fautes d'orthographe) mais **difficile** et **coûteuse** dans d'autres cas (valeur admissible mais incorrecte)

Complétude (Completeness) (1)

[Batini 06] Complétude (d'un assemblage de données): la couverture avec laquelle le phénomène observé est représenté dans l'assemblage des données

On distingue

- Complétude d'un enregistrement: présence de valeurs nulles dans un enregistrement
- Complétude d'un attribut: nombre des valeurs non nulles de l'attribut
- Complétude d'une relation: nombre des valeurs non nulles dans la relation

Aussi

- Complétude des données relationnelles (BD)
 - Complétude au niveau d'un champ ou attribut
 - Complétude au niveau d'un enregistrement (tuple)
- Complétude des données web
 - Complétude traditionnelle
 - Complétude dynamique (temporelle): complétude (degré de complétude actuelle des données) / complétabilité (décrit la manière avec laquelle ce degré évolue dans le temps)

Complétude (Completeness) (2)

[Pipino et al. 02]: 3 niveaux de complétude

- Complétude au niveau du schéma: le degré avec lequel toutes les entités et les attributs sont représentés
- Complétude au niveau de la colonne: valeurs manquantes dans 1 colonne
- Complétude au niveau de la population:
 - Si une colonne doit représenter au moins 1 occurrence de 50 départements, et qu'on n'en retrouve que 43, on parle d'incomplétude de population.

[Berti Equille 07]: définit la complétude dans le modèle relationnel

- Présence ou absence de valeurs nulles
- L'interprétation dépend du paradigme adopté: hypothèse du monde fermé/ouvert
 - Les valeurs nulles représentent des données/objets existants mais inconnus
 - Les valeurs nulles représentent des données/objets inexistantes
 - Les valeurs nulles représentent des données/objets qui pourraient exister, cependant nul ne peut affirmer leur (in)existence

Complétude (Completeness) (3)

– Hypothèse du monde fermé

- **Complétude horizontale**: le nombre de valeurs nulles dans un enregistrement ou un ensemble d'enregistrements
- **Complétude verticale**: le nombre de valeurs nulles dans le domaine de définition d'un attribut spécifique

☹ La complétude horizontale n'est pas toujours évidente à détecter

– Hypothèse du monde ouvert: la connaissance du monde réel est incomplète => ce qui n'est pas représenté (valeurs nulles) est considéré inconnu plutôt que faux, contrairement au monde fermé.

- **Couverture** de la source S : mesure le nombre d'enregistrements fournis par S , relativement à une relation universelle U
- **Densité**: mesure le volume moyen de données fournies par une source S pour un enregistrement e et un attribut a donnés

Dimensions temporelles

- Promptitude (Timeliness)
- Actualité (Currency)
- Récence (Recency)
- Volatilité (Volatility)
- Fraîcheur (Freshness)
- Obsolescence (Obsolescence)

Dimensions temporelles

Promptitude[1] (Timeliness)

[Berti-Equille 07] : décrit l'âge des données

[Wand et Wang 96]: c'est la propriété *intrinsèque* d'une donnée, et peut être mesurée avec le **retard temporaire** entre un évènement du monde réel et son enregistrement dans le système informatique

≈ [Wang et Strong 96]: désigne principalement le caractère courant ou à jour des données au moment de leur diffusion, en mesurant **l'écart entre la fin de la période de référence à laquelle les données se rapportent et la date à laquelle les données deviennent accessibles** aux utilisateurs.

[Pipino et al. 02]: le degré d'actualité des données respectivement aux tâches dans lesquelles elles sont utilisées

– Maximum $\{0, (1 - \text{currency}/\text{volatility})\}$

[Naumann et al. 99]: c'est la fréquence de MAJ des données ou la fréquence de création de nouvelles données dans une source S => c'est l'intervalle de temps séparant la date de MAJ de la donnée de la date de son utilisation (via les requêtes)

Dimensions temporelles

Actualité (Currency)

Confondue dans [Segev et al. 90] avec la notion de fraîcheur de données et définit le degré d'ancienneté des données respectivement à leurs sources

- actualité = date de livraison des données aux users – date d'extraction des données depuis les sources

[Pipino et al. 02]

- âge + (date de livraison – date de création des données (imputation des données dans le système))

[Scannapieco et al. 05]

- mesure l'obsolescence des données
- mesure la **rapidité de la mise à jour des données**

[Akoka et al. 07]: dans le domaine des entrepôts de données et données multi-sources:

- décrit le degré d'ancienneté (obsolescence) des données: délai entre la date d'extraction des données à partir des sources et sa livraison aux utilisateurs

Dimensions temporelles

Volatilité (Volatility)

[Scannapieco et al. 05]: la fréquence avec laquelle les données varient dans le temps

[Pipino et al. 02]: la durée pendant laquelle les données demeurent **valides**

- Facteur intervenant dans l'évaluation de la promptitude (timeliness) plutôt qu'une dimension à part entière

Dimensions temporelles

Fraîcheur (Freshness)

[Segev et al. 90] : anciennement définie comme étant l'actualité

[Akoka et al. 07]: est relative à l'âge des données: sont-elles suffisamment fraîches respectivement aux attentes des utilisateurs?

– la fraîcheur représente une famille de dimensions/facteurs qualité, chacun représentant un aspect de la fraîcheur, les dimensions les plus importantes:

- Actualité (Currency)
- Promptitude (Timeliness)

[QUADRIS 06]: dans le contexte des données multi-sources:

– Intervalle entre les dates d'extraction, de mise à jour et d'intégration

Dimensions temporelles

Récence (Recency)

[Peralta 06] décrit l'âge des données et définit la fraîcheur

[Shin 2003] confondue avec l'actualité des données

Dimensions temporelles Obsolescence (obsolescence)

[Gal et al. 99]: dans les systèmes de requête

Le nombre d'insertions, suppressions et modifications depuis la date de matérialisation des données

[Gancarski et al. 03]: dans les systèmes de répliquions

Appelée « ordre » et mesure le nombre de transactions de rafraîchissement ayant été commitées dans le nœud maître mais n'ayant pas été propagées aux nœuds esclaves

[Peralta 06]: dans le contexte des données multisources

Mesure le nombre de MAJ de la source de la donnée depuis son extraction

Cohérence (Consistency)

[Redman 96]

La satisfaction des contraintes d'intégrité.

Exemple: la cohérence entre l'âge d'un employé et son année de naissance, unicité de l'attribut clé, présence de doublons, etc.

C'est aussi un moyen pour mesurer l'exactitude sémantique: 2 valeurs sont incohérentes quand elles ne peuvent pas être correctes ensemble.

→ La définition des règles de cohérence aide à identifier les valeurs incorrectes.

[Pipino et al. 02]

C'est la cohérence de représentation de la même entité réelle dans des tables différentes (contrainte d'intégrité référentielle de Codd)

→ Concerne tout ce qui se rattache à la violation des règles sémantiques.

1- Violation des contraintes d'intégrités (intra et inter relation)

2- Les « data edits » ou encore (les règles logiques – concerne les données non relationnelles)

Dimensions de confiance

Crédibilité (Believability)

Objectivité (Objectivity)

Dimensions de confiance

Crédibilité (Believability)

[Wang et al. 92]

Fortement interprétable

Jugement de l'analyste nécessitant les informations suivantes

Quand les données sont-elles générées? D'où viennent-elles? Comment sont-elles obtenues? Par quels moyens sont-elles stockées dans la base?

[Pipino et al. 02]

Degré de confiance et d'exactitude des données, mesure objective donnée par une moyenne pondérée des mesures individuelles

Reflète aussi le niveau de confiance que l'on accorde aux données étant données leurs sources, mesure subjective déduite de l'expérience des experts

[Gackowski 06]

Confondue avec « Credibility »

Dimensions de confiance

Objectivité (Objectivity)

[Gackowski 06]

Données non biaisées

[Krol 08]

Il s'agit de données réelles, impartiales, basées sur des faits

Dimensions de confiance mais aussi...

On retrouve aussi

Fiabilité (Reliability)

Réputation (Reputation)

mais plutôt pour caractériser les sources des données et non la qualité intrinsèque de la donnée

Précision (Precision)

[Wand et Wang 96]

mesure le degré d'ambiguïté des données

[Redman 96]

Présente le niveau de détail de la représentation de l'information

[Peralta 06]

Métrique évaluant l'exactitude (accuracy) de la donnée (la valeur est-elle assez précise? valide? correcte?)

Pertinence (Relevancy/Relevance)

[Bovee et al. 01]

Applicabilité/adéquation à 1 domaine spécifique

[Pipino et al. 02], [Otto et Ebner 10]

Le niveau d'adéquation/utilisabilité des données à la tâche en question
C'est une dimension importante surtout dans les systèmes basés sur le web

Exemple: l'information sur les clients est souvent confrontée à un grand volume d'informations potentiellement pertinentes

Unicité (Uniqueness)

[Mecella et al. 02]

Exprime le fait que 2 ou plusieurs valeurs ne créent pas de conflits les unes par rapport aux autres

[Akoka 07], [Berti-Equille 07]

Absence de doublons dans la base

Conclusions

- La qualité des données
 - Absence de standard définissant les dimensions → objectif de notre groupe de travail
 - Les dimensions ne sont pas indépendantes les unes des autres mais plutôt corrélés
 - Si l'on favorise une dimension par rapport à d'autres lorsqu'elle nous semble la plus importante pour une application donnée, ceci peut engendrer des conséquences négatives sur les autres dimensions (par exemple exactitude VS promptitude ou cohérence VS complétude)
- La QoD, mais aussi
 - La QoS: qualité des sources
 - La QoP: qualité des processus
 - La DoM: qualité des modèles

Bibliographie (1)

[1] http://ocaq.qc.ca/terminologie/affichage_bulletin.asp?ID=112

[Akoka et al. 07] <http://www.lamsade.dauphine.fr/~goasdoue/Publications/2007/QUADRIS-ICEIS.pdf>

[Batini 06]

http://www.google.fr/url?sa=t&source=web&ct=res&cd=5&ved=0CCUQFjAE&url=http%3A%2F%2Fwww.mmsp.gov.ma%2Ffrancais%2FManifestations_fr%2FSeminaires_fr%2FDoc_Seminaires_fr%2FAdministrationElect%2FQualitedesdonnees1.pps&rct=j&q=dimensions+qualit%2E9+donn%2E9es&ei=ZHMzS6nNAoJ4ga2l_ylAg&usq=AFQjCNEYT53hull3TAGCwrYRuCysRq064A&sig2=7ZCJPn6O9BgxmwLFncflog

[Batini et Scannapieco 06] Batini C. , Scannapieco M.. "Data quality: concepts, methodologies and techniques". 2006

[Redman 96] Redman, T.: "Data Quality for the Information Age". Artech House, 1996.

[Naumann et al. 99] Naumann, F., Leser, U., Freytag, J.C.. "Quality-driven Integration of Heterogeneous Information Systems". In Proc. of the 25th Int. Conf. on Very Large Databases (VLDB'99), Scotland, 1999.

[Peralta 06] Peralta V. "Data Quality Evaluation in Data Integration Systems". PhD Thesis.2006

[Wang et Strong 96] Wang R., Strong D. "Beyond accuracy: what data quality means to data consumers". Journal of management information systems. 1996

[Pipino et al. 02] Pipino L., Lee Y.W., Wang R.Y.. "Data Quality Assessment". 2002

[Berti Equille 07] Berti Equille L.. "Quality awareness for managing and mining data". 2007

[Scannapieco et al. 05] Scannapieco M., Missier P., Batini C.. "Data quality at a glance". 2005

[Segev et al. 90] Segev, A., Weiping, F., "Currency-Based Updates to Distributed Materialized Views". In proceedings of *ICDE'90*, 1990.

[Wand et Wang 96] Wand Y., Wang R.Y.. "Anchoring Data Quality Dimensions in Ontological Foundations". Communication of the ACM, vol. 39, no. 11, 1996.

[Quadris 06] "Quality of data in multi-source information systems".

http://paristic.loria.fr/content/masse_de_donnees/talks/QUADRIS.pdf. 2005-2008.

Bibliographie (2)

- [Gal 1999] Gal, A.: "Obsolescent materialized views in query processing of enterprise information systems". In Proc. of the 1999 ACM Int. Conf. on Information and Knowledge Management (CIKM'99), pages 367-374, Kansas City, USA. 1999.
- [Gancarski+2003] Gancarski, S., Le Pape, C., Valduriez, P.: "Relaxing Freshness to Improve Load Balancing in a Cluster of Autonomous Replicated Databases". In Proc. of the 5th workshop on Distributed Data and Structures (WDAS), Thessaloniki, Greece. 2003.
- [Wang et al. 92] Wang R.Y., Reddy M.P., Kon B.H.. "Toward Quality Data: An Attribute-Based Approach". 1992.
- [Gackowski 06] Gackowski J.Z.. "Diagnostic and Functional Dependencies of Credibility". Informing Science Journal Volume 9, 2006.
- [Otto et Ebner 10] Otto B., Ebner V.. "Measuring Master Data Quality" .2010
- [Bovee et al. 01] Bovee M., Srivastava R.P., Mak B.R.. "A Conceptual Framework and Belief-Function Approach to Assessing Overall Information Quality". Proceedings of the 6th International Conference on Information Quality. 2001.
- [Mecella et al. 02] [Mecella+2002] Mecella, M., Scannapieco, M., Virgillito, A., Baldoni, R., Catarci, T., Batini, C.. "Managing Data Quality in Cooperative Information Systems". DOA. 2002.
- [Wang et Strong 96] Wang R.Y., Strong D.. "Beyond accuracy: what data quality means to data consumers". Journal of management information systems. 1996.