Calidad de datos

Claudio Gutierrez
DCC, Universidad de Chile, IMFD
2022

Calidad

calidad[1] Del lat. *qualĭtas, -ātis,* y este calco del gr. ποιότης *poiótēs.*

1. f. Propiedad o conjunto de propiedades inherentes a algo, que permiten juzgar su valor. *Esta tela es de buena calidad*.
2. f. Buena calidad, superioridad o excelencia. *La calidad de ese aceite ha conquistado los mercados*.
3. f. Adecuación de un producto o servicio a las características especificadas. *Control de la calidad de un producto.*

Dato

cri cri cri …..

Calidad

"the degree to which a set of inherent characteristics fulfill the requirements"
(General Administration of Quality Supervision, 2008);

"fitness for use" (Wang & Strong, 1996);

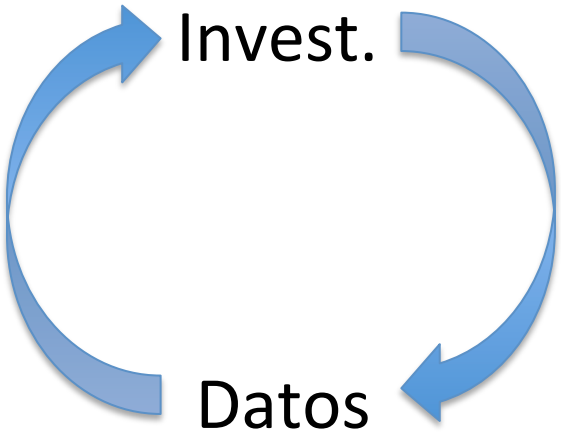
"conformance to requirements" (Crosby, 1988)

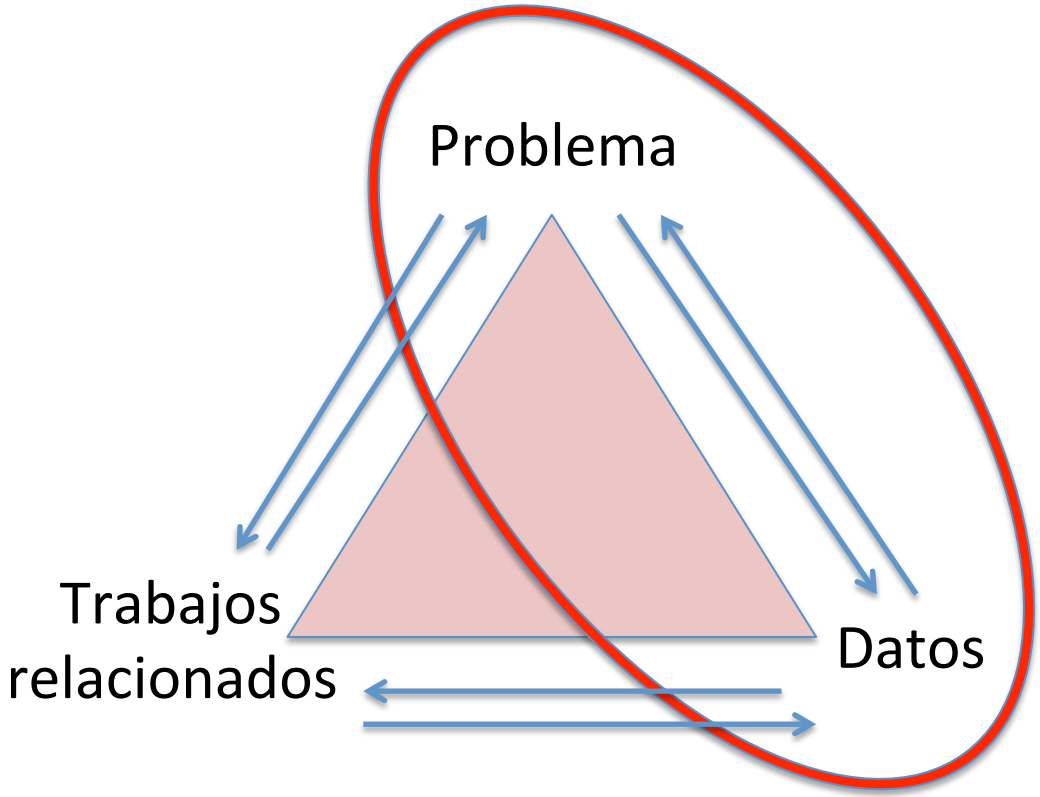
Dos enfoques

1. Investigación → Datos
   - La investigación determina los datos (necesarios)
   - Se plantea el problema, y de allí se ve qué datos son necesarios para abordarlo

2. Datos → Investigación
   - Recolectar datos y luego ver para qué investigación pueden servir
   - De acuerdo a los datos, se detemina el problema

Realmente uno solo
(con diferentes énfasis)





Los dos pilares de la calidad de datos

1. Correspondencia de los datos con el problema tratado

2. Aspectos técnicos de los datos mismos

Ambos son igualmente relevantes a la hora de abordar el problema. Usualmente (ii) ha sido subestimado. En la era de Big Data es un gran tema.

Correspondencia

Ajuste, adecuación, entre de los datos con el problema tratado:

1. Relevantes para el problema
2. Representativos del problema
3. Estadísticamente significativos
4. Con sesgo controlado
5. etc.

Nota: Se evalúan aspectos conceptuales a partir del problema definido

Aspectos técnicos de los datos

Los datos que se definieron a partir del problema:

1. Existen
2. Son accesibles
3. Están disponibles legalmente
4. Están en formatos adecuados
5. Están suficientemente “limpios”
6. Tiene documentación
7. etc.

El problema de datos (de información realmente!) de una organización o empresa

Ambos extremos (problema y datos) se mezclan. Se necesitan siempre datos confiables. No es evidente en qué se usarán. Pero si aparece un problema, se necesitan buenos datos (y a tiempo) para abordarlo.

Esto lleva al problema de la calidad de datos al interior de una organización o empresa.

“data quality depends not only on its own features but also on the business environment using the data, including business processes and business users. Only the data that conform to the relevant uses and meet requirements can be considered qualified (or good quality) data.” (Cai, Zhu, 2015)

Nuestro problema

En este curso el problema es ligeramente diferente. Nosotros queremos resolver un problema. Y para ello buscamos los datos y entramos en el ciclo problema/datos/… etc.

Luego el problema de la correspondencia es muy relevante

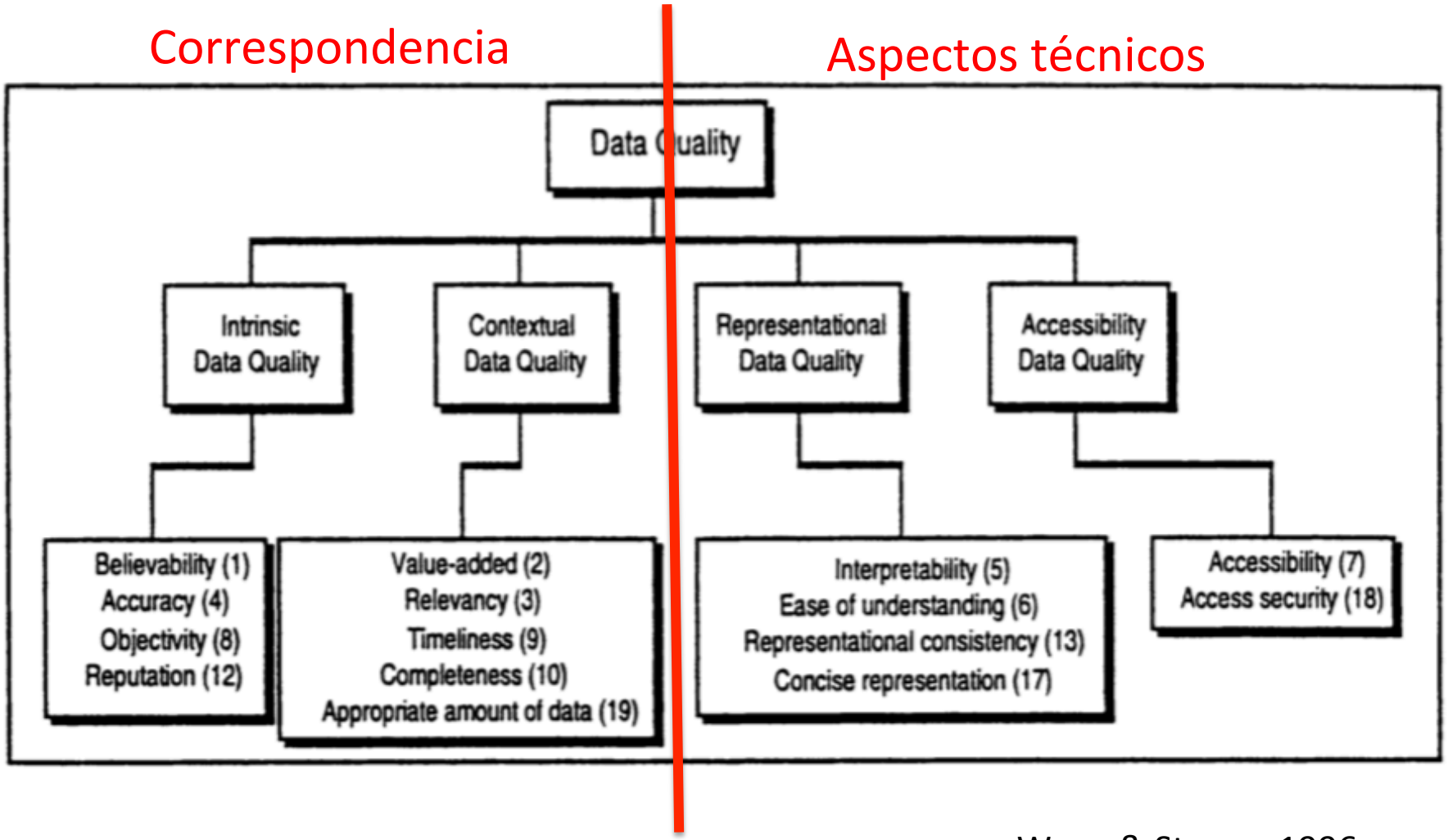
Tres tipos de "ejercicios"

1. Evaluar la calidad de datos de una organización (e.g. del lugar donde trabajan)
2. Evaluar la calidad de los datos a usar en una investigación o problema dado
3. Evaluar la calidad de un *dataset* cualquiera (e.g. entrar a https://datos.gob.cl/ y tomar un dataset)

Tres tipos de "ejercicios"

1. Evaluar la calidad de datos de una organización (e.g. del lugar donde trabajan)
2. Evaluar la calidad de los datos a usar en una investigación o problema dado
3. Evaluar la calidad de un *dataset* cualquiera (e.g. entrar a https://datos.gob.cl/ y tomar un dataset)

Data Quality (Wang & Strong 1996)



Wang & Strong, 1996

Calidad en tiempos de big data

La era de big data trae varios desafíos nuevos: gigantes volúmenes, dinamicidad, variedad (heterogeneidad), integración, etc.

Pero además, trae un desafío nuevo: los datos por sí mismo comienzan a tener valor (antes de un problema determinado). Entonces lo que mueve a conseguir datos son áreas temáticas, familias de problemas, para los cuales hay que construir fuentes de datos que tengan valor

Los desafíos del big data

The diversity of data sources brings abundant data types and complex data structures and increases the difficulty of data integration.

Data volume is tremendous, and it is difficult to judge data quality within a reasonable amount of time.

Data change very fast and the “timeliness” of data is very short, which necessitates higher requirements for processing technology.

No unified and approved data quality standards have been formed in China and abroad, and research on the data quality of big data has just begun.

(Cai & Zhu, 2015)

Calidad de datos (Cai & Zhu, 2015)

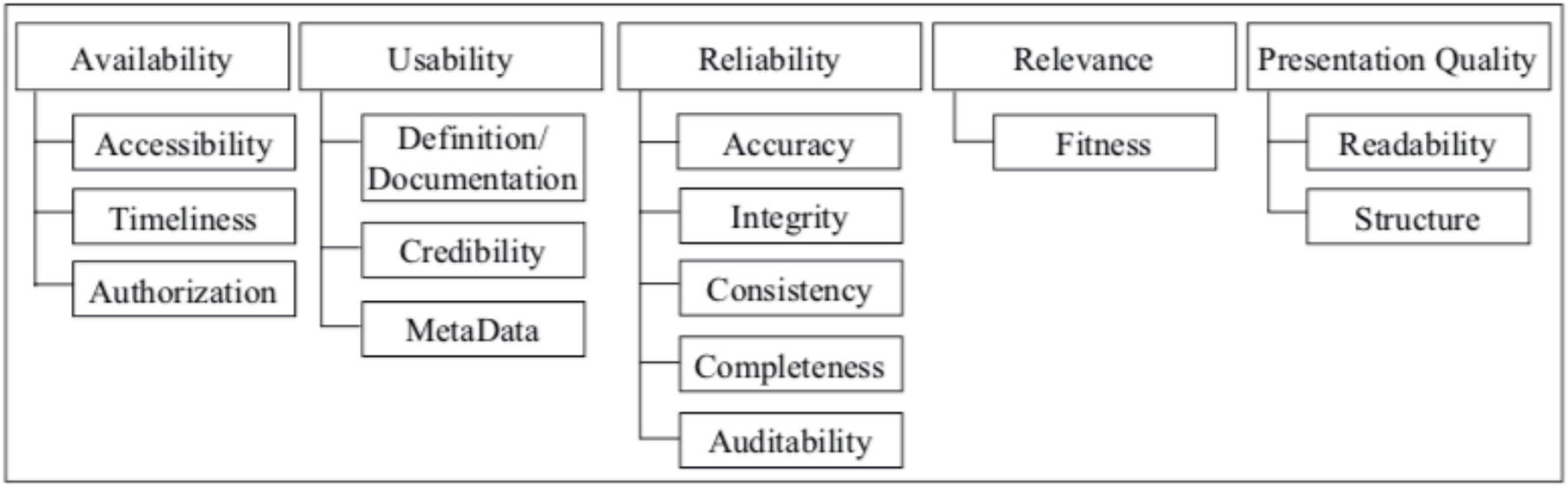


Figure 2: A universal, two-layer big data quality standard for assessment.

Dimensions	Elements	Indicators
1) Availability	1) Accessibility	▪ Whether a data access interface is provided ▪ Data can be easily made public or easy to purchase
	2) Timeliness	▪ Within a given time, whether the data arrive on time ▪ Whether data are regularly updated ▪ Whether the time interval from data collection and processing to release meets requirements
2) Usability	1) Credibility	▪ Data come from specialized organizations of a country, field, or industry ▪ Experts or specialists regularly audit and check the correctness of the data content ▪ Data exist in the range of known or acceptable values
3) Reliability	1) Accuracy	▪ Data provided are accurate ▪ Data representation (or value) well reflects the true state of the source information ▪ Information (data) representation will not cause ambiguity
	2) Consistency	▪ After data have been processed, their concepts, value domains, and formats still match as before processing ▪ During a certain time, data remain consistent and verifiable ▪ Data and the data from other data sources are consistent or verifiable
	3) Integrity	▪ Data format is clear and meets the criteria ▪ Data are consistent with structural integrity ▪ Data are consistent with content integrity
	4) Completeness	▪ Whether the deficiency of a component will impact use of the data for data with multi-components ▪ Whether the deficiency of a component will impact data accuracy and integrity
4) Relevance	1) Fitness	▪ The data collected do not completely match the theme, but they expound one aspect ▪ Most datasets retrieved are within the retrieval theme users need ▪ Information theme provides matches with users' retrieval theme
5) Presentation Quality	1) Readability	▪ Data (content, format, etc.) are clear and understandable ▪ It is easy to judge that the data provided meet needs ▪ Data description, classification, and coding content satisfy specification and are easy to understand

Table 1: The hierarchical big data quality assessment framework (partial content).

1) **Availability / Accessibility**

Whether a data access interface is provided
Data can be easily made public or easy to purchase

2) **Availability / Timeliness**

Within a given time, whether the data arrive on time
Whether data are regularly updated
Whether the time interval from data collection and processing to release meets requirements

3) Usability / Credibility

Data come from specialized organizations of a country, field, or industry

Experts or specialists regularly audit and check the correctness of the data content

Data exist in the range of known or acceptable values

1) Accuracy:

Data provided are accurate
Data representation (or value) well reflects the true state of the source information
Information (data) representation will not cause ambiguity

2) Consistency

After data have been processed, their concepts, value domains, and formats still match as before processing

During a certain time, data remain consistent and verifiable
Data and the data from other data sources are consistent or verifiable

3) Integrity

Data format is clear and meets the criteria

Data are consistent with structural integrity

Data are consistent with content integrity

4) Completeness

Whether the deficiency of a component will impact use of the data for data with multi-components

Whether the deficiency of a component will impact data accuracy and integrity

Relevance / Fitness

- The data collected do not completely match the theme, but they expound one aspect
- Most datasets retrieved are within the retrieval theme users need Information theme provides matches with users' retrieval theme

Presentation Quality / Readability

- Data (content, format, etc.) are clear and understandable
- It is easy to judge that the data provided meet needs
- Data description, classification, and coding content satisfy specifica-

TABLE 2. Data journal peer review guidelines. (Note: these are drawn from the associated websites. Some edits have been made for presentation.)

Earth System Science Data www.earth-system-science-data.net/review/ms_evaluation_criteria.html	*Geoscience Data Journal* http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%292049-6060/homepage/guidelines_for_reviewers.htm	*Scientific Data* www.nature.com/scientificdata/guide-to-referees/
I. Read the manuscript: 1. Are the data and methods presented new? 2. Is there any potential of the data being useful in the future? 3. Are methods and materials described in sufficient detail? 4. Are any references/citations to other datasets or articles missing or inappropriate? **II. Check the data quality:** 5. Is the dataset accessible via the given identifier? 6. Is the dataset complete? 7. Are error estimates and sources of errors given (and discussed in the article)? 8. Is the accuracy, calibration, processing etc. state of the art? 9. Are common standards used for comparison? **III. Consider article and dataset:** 10. Are there any inconsistencies within these, implausible assertions or data or noticeable problems which would suggest the data are in error (or worse). 11. If possible, apply tests (e.g., statistics). 12. Unusual formats or other circumstances which impede such tests as are usual in your discipline may raise suspicion. **IV. Check the presentation quality:** 13. Is the dataset usable in its current format and size? 14. Is the formal metadata appropriate? **Finally:** By reading the article and downloading the dataset would you be able to understand and (re-)use the dataset in the future?	**I. Data description document** 1. Is the method used to create the data of a high scientific standard? 2. Is enough information provided (in metadata also) to enable the data to be re-used or the experiment to be repeated? 3. Does the document provide a comprehensive description of all the data that is there? 4. Does the data make an important and unique contribution to the geosciences? 5. What range of applications to geosciences does it have? 6. Are all contributors and existing work acknowledged? 7. Does the Data Paper contain sufficient citation information of the dataset, e.g., dataset DOI, name of data center etc.? **II. Metadata** 8. Does the metadata establish the ownership of the data fairly? 9. Is enough information provided (in data description document also) to enable the data to be re-used or the experiment to be repeated? 10. Are the data present as described, and accessible from a registered repository using the software provided? **III. The data themselves** 11. Are the data easily readable, e.g. do they use standard or community formats? 12. Are the data of high quality e.g., are error limits and quality statements adequate to assess fitness for purpose, is spatial or temporal coverage good enough to make the data useable? 13. Are the data values physically possible and plausible? 14. Are there missing data that might compromise its usefulness?	**I. Experimental Rigor and Technical Data Quality** 1. Were the data produced in a rigorous and methodologically sound manner? 2. Was the technical quality of the data supported convincingly with technical validation experiments and statistical analyses of data quality or error, as needed? 3. Are the depth, coverage, size, and/or completeness of these data sufficient for the types of applications or research questions outlined by the authors? **II. Completeness of the Description** 4. Are the methods and any data processing steps described in sufficient detail to allow others to reproduce these steps? 5. Did the authors provide all of the information needed for others to reuse this dataset, or integrate it with other data? 6. Is this Data Descriptor, in combination with any repository metadata, consistent with relevant minimum information or reporting standards? **III. Integrity of the Data Files and Repository Record** 7. Have you confirmed that the data files deposited by the authors are complete and match the descriptions in the Data Descriptor? 8. Have these data files been deposited in the most appropriate available data repository?

(Mayernik et al. 2015)

Evaluación de datasets

I. **Data description document**
II. **Metadata**
III. **The data themselves**

(Geoscience Data Journal; en: Mayernik et al. 2015)

I. Data description document

1. Is the method used to create the data of a high scientific standard?
2. Is enough information provided (in metadata also) to enable the data to be re-used or the experiment to be repeated?
3. Does the document provide a comprehensive description of all the data that is there?
4. Does the data make an important and unique contribution to the geosciences?
5. What range of applications to geosciences does it have?
6. Are all contributors and existing work acknowledged?
7. Does the Data Paper contain sufficient citation information of the dataset, e.g., dataset DOI, name of data center etc.?

II. Metadata

8. Does the metadata establish the ownership of the data fairly?
9. Is enough information provided (in data description document also) to enable the data to be re-used or the experiment to be repeated?
10. Are the data present as described, and accessible from a registered repository using the software provided?

III. The data themselves

11. Are the data easily readable, e.g. do they use standard or community formats?

12. Are the data of high quality e.g., are error limits and quality statements adequate to assess fitness for purpose, is spatial or temporal coverage good enough to make the data useable?

13. Are the data values physically possible and plausible?

14. Are there missing data that might compromise its usefulness?