Marco de trabajo para la Gestión de la Calidad de Datos en Gobierno Digital

Introducción

Uruguay Presidencia | agesic

Temas

- Motivación y objetivos del Marco de trabajo
- Principales componentes del Marco de trabajo
- Escenario de trabajo
- Marco teórico

Los temas se presentan en forma de “guía rápida”. Por más detalles consultar el documento:
“Framework para la Gestión de la Calidad de Datos en Gobierno Digital” [AGE2020].

Temas

- Motivación y objetivos del Marco de trabajo
- Principales componentes del Marco de trabajo
- Escenario de trabajo
- Marco teórico

Motivación y objetivos
Motivación

- En un contexto de gobierno digital, las organizaciones gestionan cantidades de datos cada vez más grandes
- La calidad de datos en estos escenarios es de suma importancia, dado que tiene un fuerte impacto en:
  - la calidad de los servicios públicos que se brindan a los ciudadanos
  - la definición de estrategias y políticas públicas

Motivación y objetivos
Motivación

- El término Calidad de Datos se utiliza generalmente en relación a un conjunto de características que los datos deben poseer (p. ej. precisión, completitud).
- Calidad de Datos también se ha definido como "adecuación al uso", poniendo énfasis en su naturaleza subjetiva.

[SC02]

Motivación y objetivos
Motivación

- La gestión de la calidad de datos en escenarios de gobierno digital implica varios desafíos, debido a características de estos escenarios [Boy11][Tep17]:
  - se maneja una gran cantidad de datos (p. ej. de todo un país)
  - la estructura de las bases de datos suele evolucionar por cambios en normativas vigentes
  - muchos datos tienen carácter probatorio (no pueden modificarse)
  - se apunta a garantizar que los ciudadanos suministren la misma información solo una vez a la administración pública
  - las organizaciones a menudo utilizan datos gestionados por otras, por lo que no se tiene control directo sobre la calidad de estos datos

Motivación y objetivos
Motivación

- Gestionar la calidad de datos en una organización implica distintas tareas como medir, analizar, mejorar y controlar los distintos aspectos de la calidad de los datos
- Resulta de interés entonces contar con mecanismos, guías y recomendaciones que asistan a las organizaciones en la gestión de la calidad de los datos en escenarios de gobierno digital

Motivación y objetivos

Objetivo general del Marco de trabajo

- contribuir a la sistematización de la gestión de la calidad de datos en organizaciones vinculadas a gobierno digital en Uruguay, con el fin de mejorar la calidad de los datos que se generan y/o utilizan en el Estado Uruguayo

- se apunta a que se pueda contar con evidencia de mayor calidad para la toma de decisiones estratégicas, que permitan brindar mejores servicios y productos al ciudadano de forma más eficiente

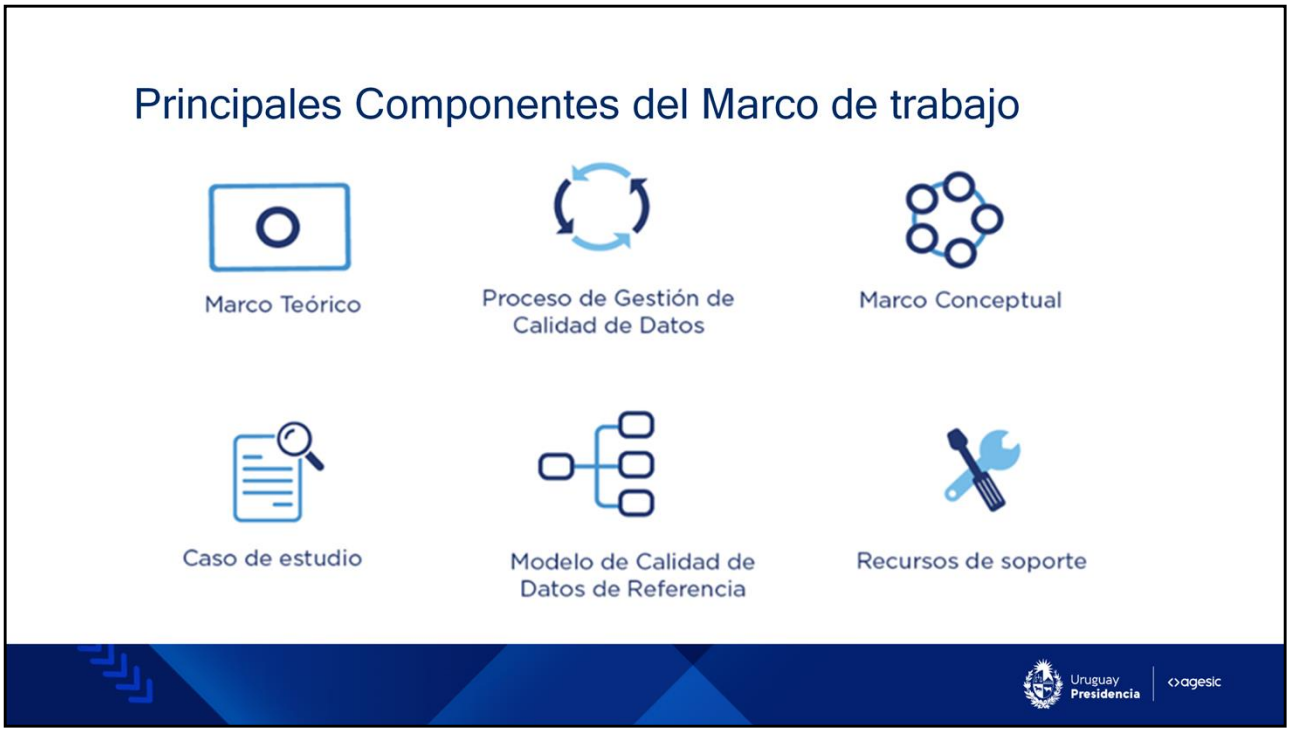
Motivación y Objetivos

Objetivos específicos del Marco de trabajo

1) proveer un marco conceptual de los distintos elementos involucrados en la gestión de calidad de datos
2) brindar un modelo de calidad de datos de referencia
3) proponer un proceso que guíe a las organizaciones en la gestión de la calidad de los datos
4) describir fundamentos teóricos relevantes
5) sugerir recursos de soporte para referencia y apoyo
6) mostrar la utilización del *marco de trabajo* en casos de estudio

Temas

- Motivación y objetivos del Marco de trabajo
- Principales componentes del Marco de trabajo
- Escenario de trabajo
- Marco teórico



El Marco Teórico proporciona fundamentos teóricos relevantes para gestionar la calidad de datos en contextos de gobierno así como para trabajar con el *marco de trabajo* propuesto. En particular el marco aborda temas de calidad de datos, modelos de calidad de datos, gestión de la calidad de datos, calidad de datos en gobierno y calidad en datos geográficos.

El Marco Conceptual define los principales conceptos relativos a la gestión de calidad de datos en contextos de gobierno, que apuntan a facilitar la comprensión de dichos contextos por parte de los distintos actores involucrados. Asimismo, estos conceptos constituyen la base conceptual sobre la cual se apoya *marco de trabajo*.

El Proceso para la Gestión de Calidad de Datos define los roles involucrados y las etapas a seguir para gestionar la calidad de datos en un escenario de gobierno. Las actividades incluyen, entre otras, la caracterización del escenario, examinar los datos objetivo y definir un modelo de calidad de datos.

El Modelo de Calidad de Datos de Referencia provee un conjunto extensible de elementos de calidad (por ejemplo: dimensiones, factores, métricas) con el fin de guiar y facilitar la definición de modelos de calidad de datos para escenarios de trabajo específicos.

Los Recursos de Soporte son herramientas, técnicas y estándares que brindan soporte y sirven de referencia para la gestión de calidad de datos así como para la aplicación del *marco de trabajo*.

El Caso de Estudio plantea un escenario de trabajo concreto para motivar, ejemplificar y guiar la gestión de la calidad de datos, en particular, cuando se utiliza el *marco de trabajo*.

Principales Componentes del Marco de trabajo

Marco Teórico

- Proporciona fundamentos teóricos relevantes para gestionar la calidad de datos en gobierno digital así como para trabajar con el *marco de trabajo*:
  - calidad de datos
  - modelos de calidad de datos
  - gestión de la calidad de datos
  - calidad de datos en gobierno digital
  - calidad en datos geográficos

Principales Componentes del Marco de trabajo

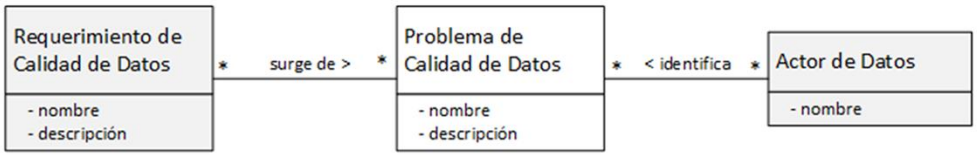
Marco Conceptual

- Define los principales conceptos relativos a la gestión de la calidad de datos en gobierno digital apuntando a:
  - facilitar la comprensión de la gestión de la calidad de datos en estos escenarios
  - brindar una base conceptual para el *marco de trabajo*
- Algunos conceptos que define el Marco Conceptual son:
  - Colección de Datos, Cliente de Datos, Entidad de Negocio
  - Actor de Datos, Requerimiento y Problema de Calidad de Datos
  - Modelo de Calidad, Técnica de Calidad de Datos, Perfil de Evaluación

Principales Componentes del Marco de trabajo

Marco Conceptual

Los problemas de Calidad de Datos refieren a cuestiones que se identifican como problemas de los datos (p. ej. direcciones no actualizadas) y que generalmente tienen consecuencias visibles para las organizaciones (p. ej. clientes no reciben facturas). Estos problemas son identificados principalmente por actores de datos.



Principales Componentes del Marco de trabajo

Modelo de Calidad de Datos de Referencia

- Uno de los elementos clave para gestionar la calidad de datos en un escenario (conjunto de sistemas, datos, requerimientos, etc.) es un Modelo de Calidad de Datos
- Un Modelo de Calidad de Datos organiza de forma jerárquica (en dimensiones, factores…) los aspectos de calidad de datos relevantes para el escenario. Por ej:
  - Exactitud
    - Correctitud Sintáctica
    - Correctitud Semántica
  - Frescura
    - Actualidad
    - Oportunidad

Principales Componentes del Marco de trabajo

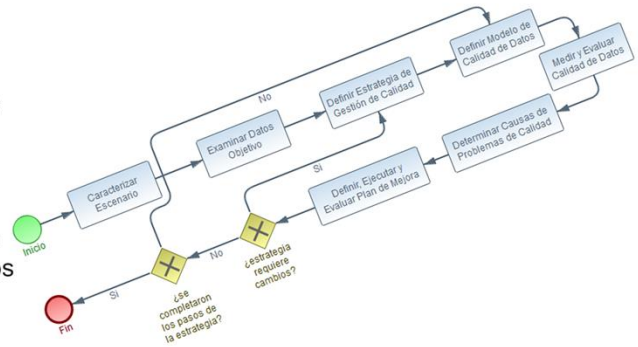
Modelo de Calidad de Datos de Referencia

- Provee un conjunto extensible e instanciable de elementos de calidad de datos (dimensiones, factores, métricas)
- Tiene como fin guiar y facilitar la definición de modelos de calidad de datos para escenarios de trabajo específicos
- El modelo incluye cinco dimensiones de calidad de datos y diecisiete factores

Principales Componentes del Marco de trabajo

Proceso de Gestión de Calidad de Datos

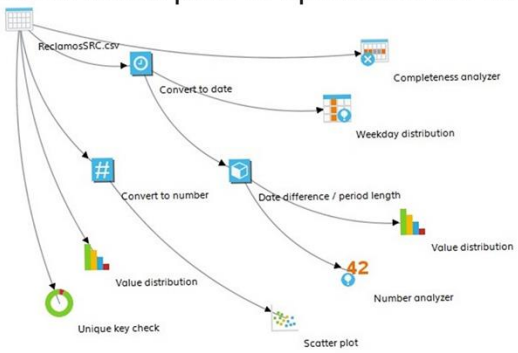
- Define los roles y etapas a seguir para gestionar la calidad de datos en un escenario de gobierno digital
  - Roles:
    - Responsable de calidad de datos
    - Analista de calidad de datos
    - ……..
  - Actividades:
    - Caracterizar escenario de trabajo
    - Definir modelo de calidad de datos
    - ……..



Principales Componentes del Marco de trabajo

Recursos de Soporte

- Conjunto extensible de artefactos que brindan soporte o sirven de referencia para la gestión de la calidad de datos así como para la aplicación del *marco de trabajo*:
  - herramientas
  - servicios
  - estándares
  - fuentes de consulta



Principales Componentes del Marco de trabajo

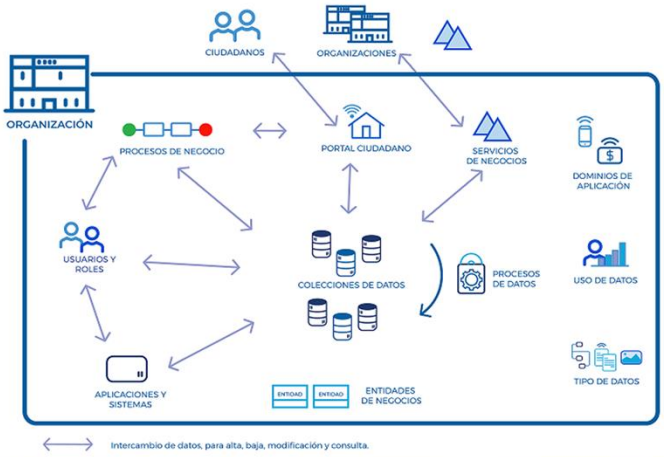
Caso de Estudio

- Plantea un escenario concreto para motivar, ejemplificar y guiar la gestión de la calidad de datos con el *marco de trabajo*
- La realidad del caso involucra:
  - varias organizaciones de gobierno
  - distintos tipos de sistemas / aplicaciones
  - distintos tipos de datos
  - distintos usos de los datos
- Está inspirado en el Sistema Único de Respuesta de la Intendencia de Montevideo

Temas

- Motivación y objetivos del Marco de trabajo
- Principales componentes del Marco de trabajo
- Escenario de trabajo
- Marco teórico

Escenario de Trabajo
Descripción General



contexto de gobierno digital donde distintas organizaciones llevan a cabo procesos de negocio y colaboran entre sí con el fin de ofrecer servicios públicos a los ciudadanos

Uruguay Presidencia | agesic

Las organizaciones se apoyan en aplicaciones y sistemas para llevar a cabo sus procesos de negocio, que pueden ser tanto implícitos como estar explícitamente implementados en sistemas especializados. Estas aplicaciones y sistemas utilizan colecciones de datos de distintos tipos (p. ej. bases de datos relacionales, archivos de texto) que almacenan la información de

las entidades de negocio relevantes para la organización (p. ej. certificados, empresas, ciudadanos, beneficios). Las entidades de negocio (p. ej. ciudadano) tienen atributos (p. ej. nombre,

dirección) que pueden ser de distinto tipo (p. ej. alfanuméricos, geográficos, imágenes).

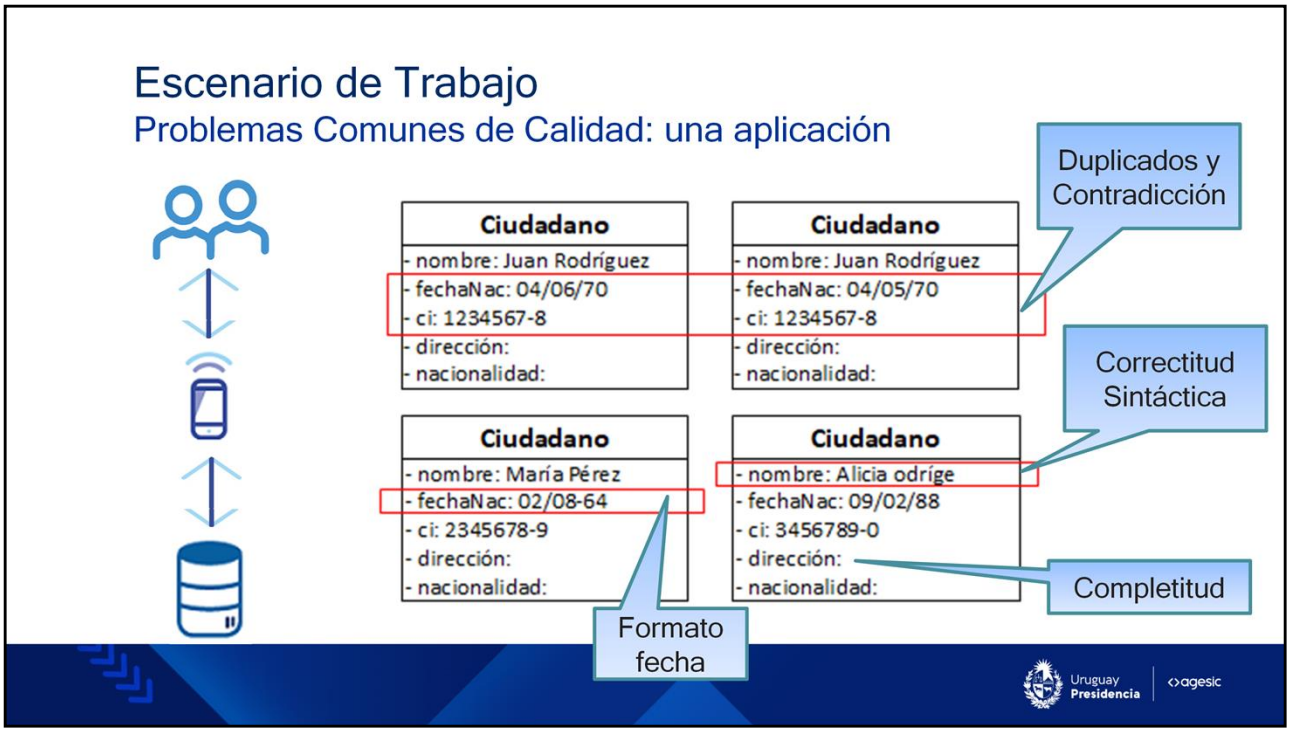
Las organizaciones ofrecen un portal a través del cual interactúan con los ciudadanos, permitiéndoles participar en procesos de negocio (p. ej. realizar trámites) así como acceder a los datos que almacenan. Los organizaciones también interactúan entre sí (p. ej. para consultar datos) a través de servicios de negocio.

Los datos en las colecciones son creados, borrados, modificados y consultados de forma manual (i.e. con intervención de una persona) o automática por distintos tipos de clientes de datos. Los clientes de datos pueden ser aplicaciones, sistemas, procesos de negocio, servicios de negocio, portal ciudadano, usuarios con distintos roles (p. ej. técnico, gerente) y procesos de datos (p. ej. procesos ETL1 para cargar un data warehouse).

Los organizaciones pueden operar en distintos dominios de aplicación (p. ej. salud, energía, telecomunicaciones, finanzas) y pueden utilizar los datos con distintos fines (p. ej. operativa

de la organización, toma de decisiones).

La Figura presenta un ejemplo de escenario de trabajo, enmarcado en el proceso de negocio que permite la generación de certificados de nacido vivo en Uruguay.



Existen varios problemas de calidad de datos que pueden surgir en el marco de una única aplicación que almacena datos en una única colección de datos (p. ej. una base de datos

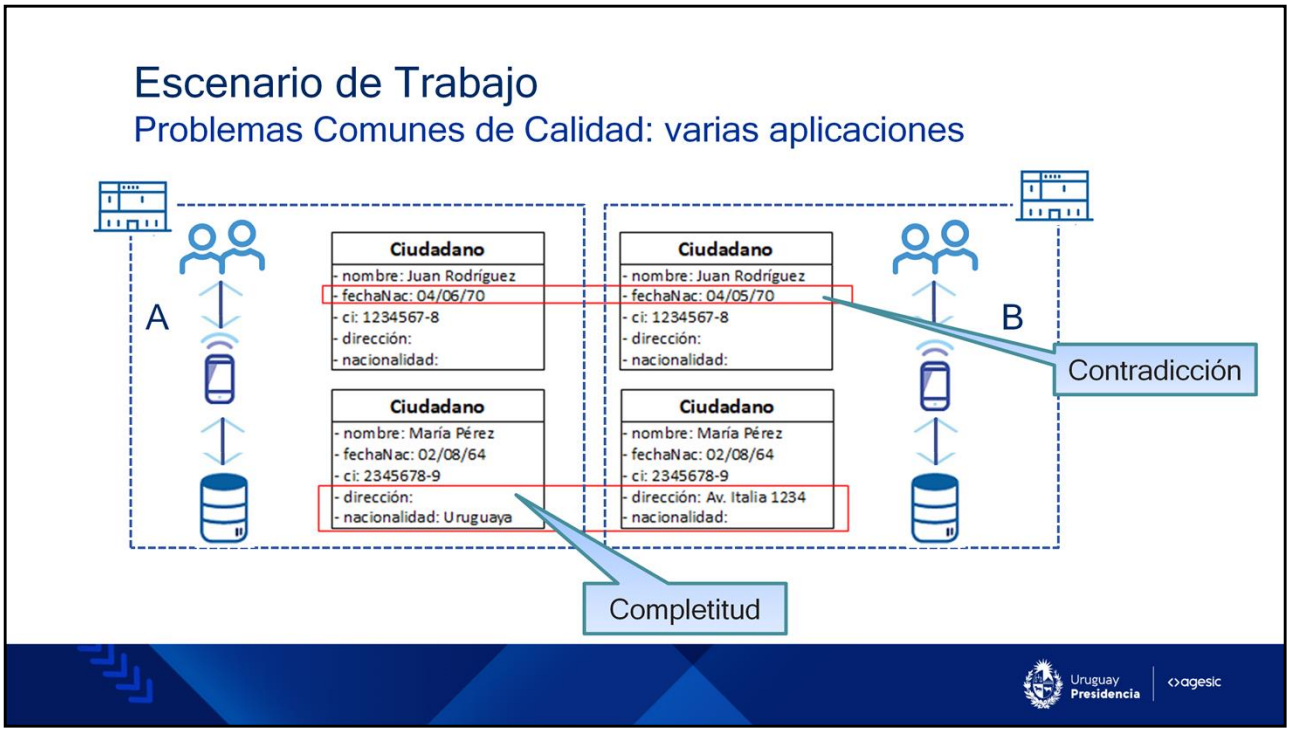
relacional). La Figura presenta algunos de estos problemas, los cuales hacen referencia al formato, unicidad y exactitud de los datos.

Con respecto al formato de los datos se puede observar que la fecha de nacimiento de la ciudadana María Pérez no sigue el formato de fecha establecido (i.e. dd/mm/aa).

En relación a la unicidad de los datos se puede observar que los datos del ciudadano Juan Rodríguez se encuentran duplicados en la colección de datos. Además, estos datos duplicados

presentan contradicciones entre sí en cuanto a la fecha de nacimiento del ciudadano.

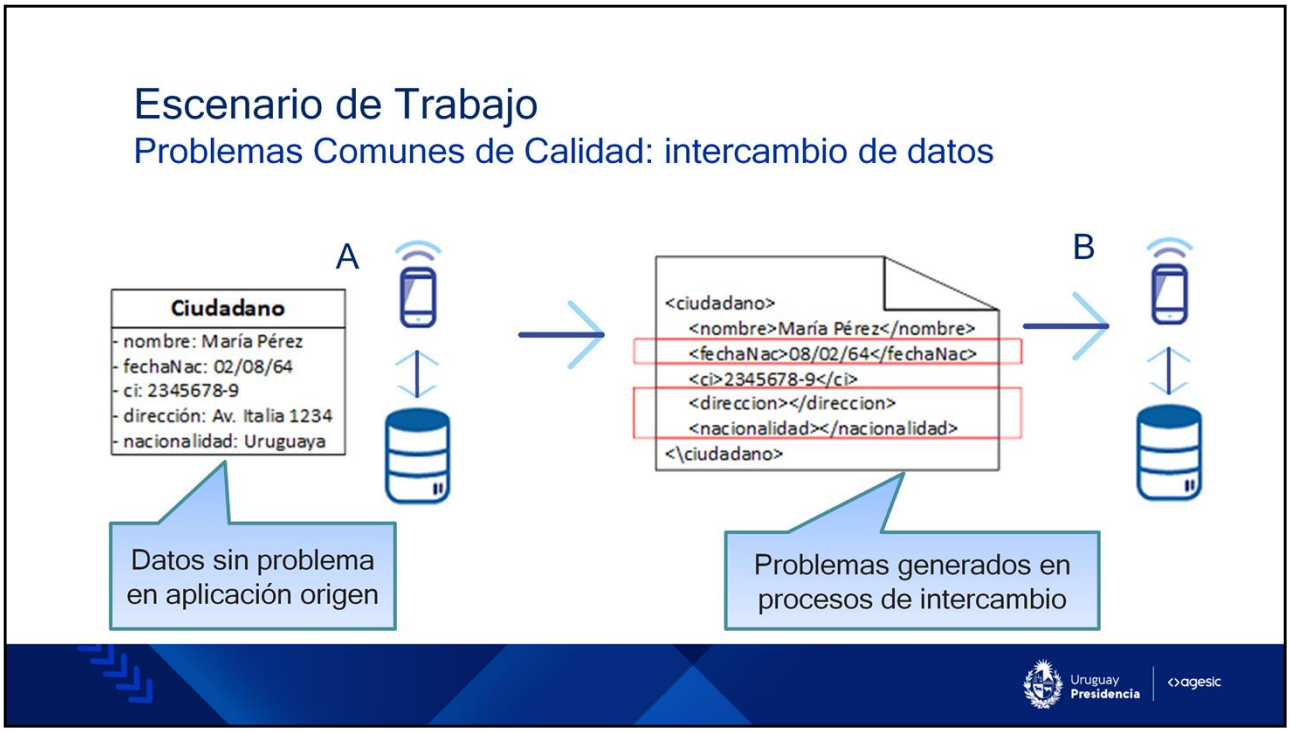
Por último, con respecto a la exactitud de los datos se puede observar que los datos del nombre de la ciudadana llamada Alicia Rodríguez no son sintácticamente correctos.



Los problemas de calidad en los datos tienden a crecer rápidamente cuando se consideran varias aplicaciones que pueden utilizar una o más colecciones de datos.

En este caso se observa cómo dos aplicaciones, que utilizan su propia colección de datos, manejan datos de un mismo ciudadano que son incompletos y contradictorios entre sí.

En particular, la fecha de nacimiento del ciudadano Juan Rodríguez que maneja la aplicación A es distinta a la que maneja la aplicación B. Asimismo, la aplicación A no tiene datos de la dirección de la ciudadana María Pérez mientras que la aplicación B sí. De forma similar, la aplicación B no tiene datos de la nacionalidad de la ciudadana María Pérez mientras que la aplicación A sí.



Aún en el caso ideal en que las colecciones de datos que manejan las aplicaciones no presenten problemas de calidad, estos problemas pueden surgir al intercambiar datos entre distintas

aplicaciones de la misma o de diferentes organizaciones.

En este caso, si bien los datos de la ciudadana María Pérez no tienen problemas de calidad en la colección de datos de la aplicación A, al intercambiarlos con la aplicación B se originan

los siguientes problemas:

- se intercambian los valores del día y mes en la fecha de nacimiento
- los datos de la dirección y nacionalidad no se incluyen en los datos intercambiados

Como consecuencia, la aplicación B recibe datos con problemas de calidad que fueron originados por los procesos de intercambio de datos.

Temas

- Motivación y objetivos del Marco de trabajo
- Principales componentes del Marco de trabajo
- Escenario de trabajo
- Marco teórico

Principales Componentes del Marco de trabajo

Marco de trabajo

Marco Teórico
Introducción

- Proporciona fundamentos teóricos relevantes para gestionar la calidad de datos en gobierno digital así como para trabajar con el marco de trabajo:
  - Calidad de Datos
  - Modelos de Calidad de Datos
  - Gestión de la Calidad de Datos
  - Calidad de Datos en Gobierno Digital
  - Calidad en Datos Geográficos

En el marco de esta guía se describen:

- Calidad de Datos
- Modelos de Calidad de Datos
- Gestión de Calidad de Datos

Marco Teórico
Calidad de Datos

- El término calidad de datos se utiliza con referencia a un conjunto de características que deben poseer los datos [SC02]:
  - correctitud, grado de actualización, etc.

- También se define como:
  - la capacidad de cumplir con los requerimientos necesarios para el uso de los datos [Ols03]
  - la adecuación de los datos para su uso (i.e. fitness for use) [WS96] [SC02] [Nee05] [Str97] [Cha05] [TB98]

Marco Teórico
Calidad de Datos: Dimensiones

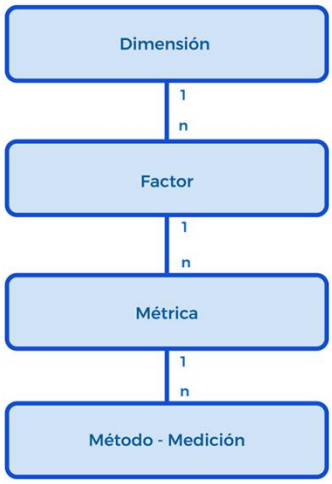
- La calidad de datos está relacionada a un conjunto de dimensiones que agrupan características de calidad [SC02].

- Las dimensiones de calidad de datos más a menudo consideradas son [Fox94]:
  - Exactitud
  - Actualidad
  - Completitud
  - Consistencia

Marco Teórico

Modelos de Calidad de Datos

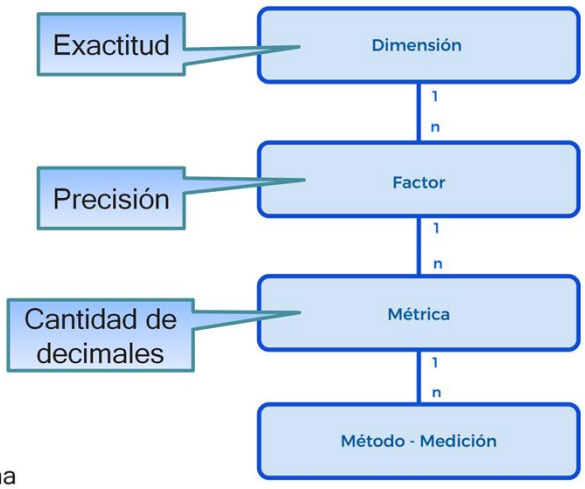
- La calidad de datos es entonces un concepto multi-facético [Etc08][Ako07]:
  - cada dimensión aborda un aspecto de la calidad
- Siguiendo este enfoque es posible definir una jerarquía de conceptos de calidad que permiten definir:
  - Modelos de Calidad de Datos



Marco Teórico

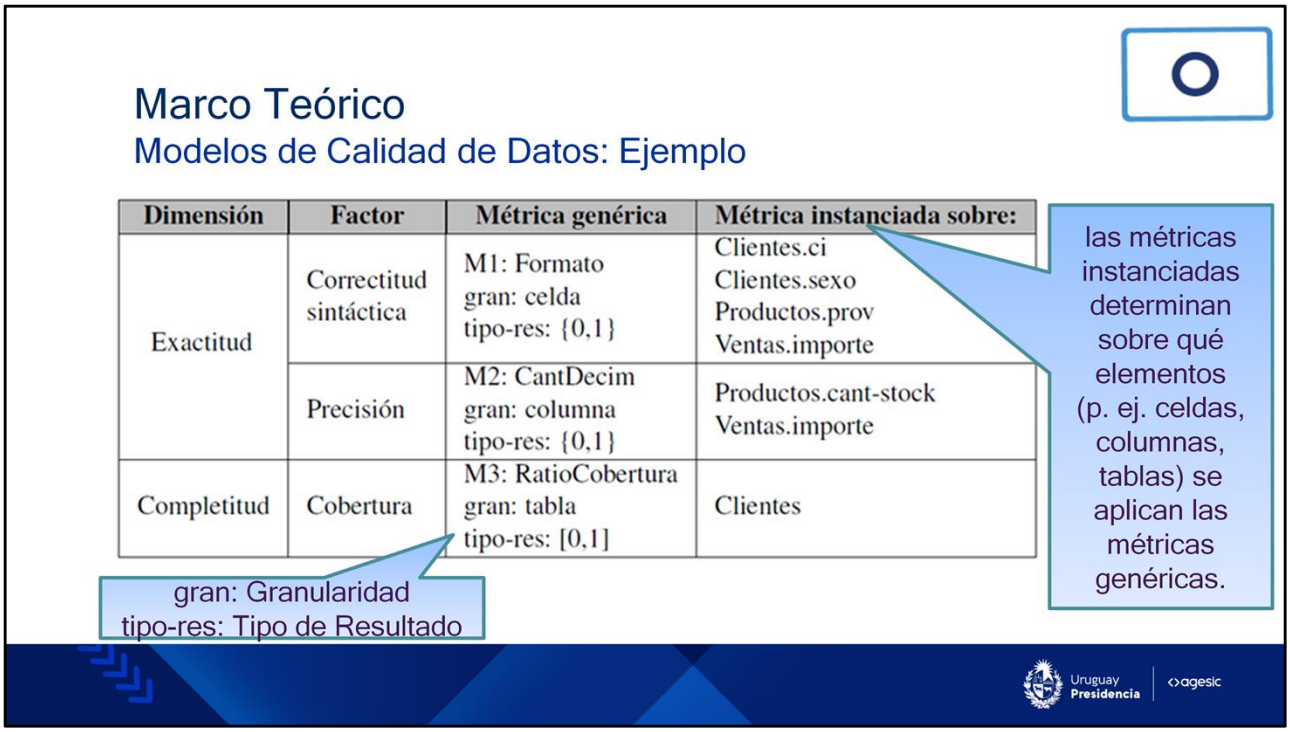
Modelos de Calidad de Datos

- Dimensión:
  - captura una faceta de la calidad de datos
  - conjunto de factores con un mismo propósito
- Factor:
  - aspecto particular de una dimensión
  - puede ser medido con distintas métricas
- Métrica
  - define la forma de medir un factor
  - puede ser medida por diferentes métodos
- Método de Medición
  - proceso que implementa una métrica
  - permite obtener medidas de calidad para la misma



Marco Teórico
Modelos de Calidad de Datos

- Es necesario definir un modelo de calidad adecuado a las necesidades y prioridades de los consumidores de los datos.
- El modelo de calidad de datos define:
  - qué dimensiones de calidad se consideran
  - sobre qué datos se aplican dichas dimensiones
  - cómo se miden
- El modelo de calidad de datos guía la gestión de la calidad para un conjunto de datos específico



Exactitud: Proximidad entre un valor de datos v y un valor de datos v', considerado como la representación correcta del fenómeno del mundo real que v intenta representar. (Adaptado de [BS16])

Correctitud Sintáctica: Proximidad entre el valor v de un atributo y los elementos del dominio de definición de dicho atributo. (Adaptado de [BS16])

Precisión: Captura el grado de detalle que posee un dato que lo hace útil para un determinado uso o que permite discriminarlo de otros datos que no son exactamente iguales. (Adaptado de [BS16]) y [ISO08])

Completitud: Captura la medida en que los datos son de la amplitud, profundidad y alcance suficientes para una determinada tarea. (Adaptado de [BS16])

Cobertura: Captura la proporción entre la cantidad de entidades existentes en una determinada colección de datos, y el total de entidades que deberían existir en dicha colección.

Marco Teórico

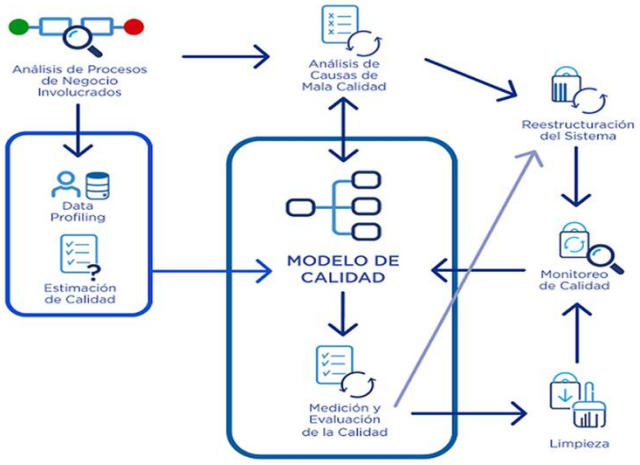
Gestión de la Calidad de Datos

- Es necesario La gestión de la calidad de datos es la tarea de medir, analizar, mejorar y controlar los distintos aspectos de calidad de los datos para un escenario específico
- Para esto se requiere contar con metodologías, técnicas, modelos, herramientas, etc

[BS16]

Marco Teórico

Gestión de la Calidad de Datos : Etapas



Motivación y objetivos
Gestión de la Calidad de Datos

- El Análisis de Procesos de Negocio Involucrados permite identificar el conjunto de datos relevantes
- Data Profiling permite conocer estos datos y realizar una primera Estimación de Calidad de los mismos
- En base a esta estimación y a un primer Análisis de las Causas de la Mala Calidad, se define el Modelo de Calidad
- Este modelo permite realizar la Medición y Evaluación de la Calidad de los datos

Motivación y objetivos
Gestión de la Calidad de Datos

- En base a los resultados de la medición y evaluación, se pueden determinar acciones para la mejora de la calidad:
  - Re-estructuración del Sistema
  - Limpieza de los Datos
- Se requiere también un Monitoreo de la Calidad constante, que puede requerir ajustes al modelo de calidad así como nuevas mediciones y evaluaciones

Referencias

[AGE2020] AGESIC. Marco de referencia para la gestión de calidad de datos. https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/comunicacion/publicaciones/marco-referencia-para-gestion-calidad-datos

[Ako07] Jacky Akoka, Laure Berti-Equille, Omar Boucelma, Mokrane Bouzeghoub, Isabelle Comyn-Wattiau, Mireille Cosquer, Virginie Goasdoué-Thion, Zoubida Kedad, Sylvaine Nugier, Verónika Peralta y col. «A Framework for Quality Evaluation in Data Integration Systems.» En: ICEIS (3). 2007, págs. 170-175.

[Boy11] Isabelle Boydens. «Strategic Issues Relating to Data Quality for E-Government: Learning from an Approach Adopted in Belgium». En: Practical Studies in EGovernment: Best Practices from Around the World. New York, NY: Springer

[BS16] Carlo Batini y Monica Scannapieco. Data and Information Quality. Springer International Publishing, 2016. DOI: 10.1007/978-3-319-24106-7.

[Cha05] Arthur D Chapman. Principles of data quality. GBIF, 2005.

Referencias

[Etc08] Lorena Etcheverry, Verónika Peralta y Mokrane Bouzeghoub. «Qbox-foundation:a metadata platform for quality measurement». En: proceeding of the 4th Workshop on Data and Knowledge Quality (QDC'2008). 2008.

[Fox94] Christopher Fox, Anany Levitin y Thomas Redman. «The notion of data and its quality dimensions». En: Information Processing & Management 30.1 (1994), págs. 9-19. ISSN: 0306-4573. DOI: http://dx.doi.org/10.1016/0306-4573(94)90020-5.

[ISO08] ISO/IEC. ISO/IEC 25012:2008 - Software engineering - Software product Quality Requirements and Evaluation (SQuaRE) - Data quality model. Estándar. International Organization for Standardization (ISO), dic. de 2008. URL: https://www.iso.org/standard/35736.html.

[Nee05] M Pamela Neely. «The product approach to data quality and fitness for use: a framework for analysis». En: Proc. of 10th International Conference on Information Quality (2005).

[Ols03] Jack E Olson. Data quality: the accuracy dimension. Morgan Kaufmann, 2003.

Referencias

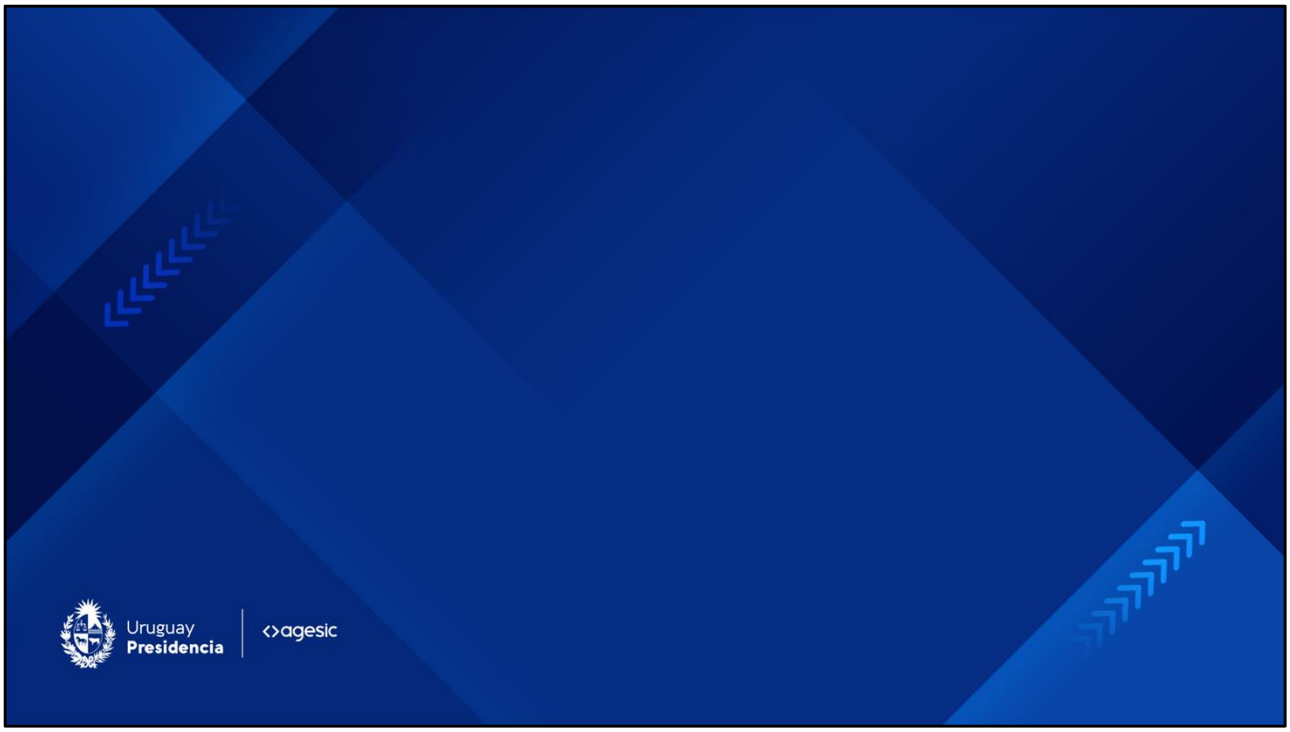
[SC02] Monica Scannapieco y Tiziana Catarci. «Data quality under a computer science perspective». En: Archivi & Computer 2 (2002), págs. 1-15.

[Str97] Diane M. Strong, Yang W. Lee y Richard Y. Wang. «Data Quality in Context». En: Commun. ACM 40.5 (mayo de 1997), págs. 103-110. ISSN: 0001-0782. DOI: 10.1145/253769.253804.

[TB98] Giri Kumar Tayi y Donald P. Ballou. «Examining Data Quality». En: Commun. ACM 41.2 (feb. de 1998), págs. 54-57. ISSN: 0001-0782. DOI: 10.1145/269012.

[Tep17] Jaak Tepandi, Mihkel Lauk, Janar Linros, Priit Raspel, Gunnar Piho, Ingrid Pappel y Dirk Draheim. «The Data Quality Framework for the Estonian Public Sector and Its Evaluation». En: Transactions on Large-Scale Data- and Knowledge-Centered Systems XXXV. Berlin, Heidelberg: Springer Berlin Heidelberg, 2017, págs. 1-26. ISBN: 978-3-662-56121-8. DOI: 10.1007/978-3-662-56121-8_1.

[WS96] Richard Y. Wang y Diane M. Strong. «Beyond Accuracy: What Data Quality Means to Data Consumers». En: J. Manage. Inf. Syst. 12.4 (mar. de 1996), págs. 5-33. ISSN: 0742-1222.


Uruguay
Presidencia
agesic