Handling Context in Data Quality Management

PhD student: Flavia Serra
fserra@fing.edu.uy

Uruguayan-French co-supervised project.
Supervisors: PhD Adriana Marotta[1], PhD Patrick Marcel[2], PhD Verónika Peralta[2]
[1] Universidad de la República
[2] Université de Tours

Outline

- Introduction & Motivation
- Systematic Literature Review
- The PhD Project
- Planning

INTRODUCTION & MOTIVATION

Introduction & Motivation

- Data Quality (DQ) is defined as **fitness for use**:
  - data could be adequate for some use,
  - but inadequate for other uses

→ **DQ depends on the context**

- DQ is recognized to be multidimensional [1]. This means that a set of DQ dimensions:
  - accuracy, completeness, consistency, uniqueness, etc.
  - express the characteristics that data should have according to their use.

- Frequently, data do not verify these characteristics or they verify them at different degrees.
  - Data Quality Problems

Introduction & Motivation

- Example: A relation *Movies* with Data Quality Problems [2].
  Domain rule: $\forall t \in R$: (t.lastYearRemake > t.year) gives **context** to *Movies* table

swapped names

inconsistent

Code	Title	Director	Year	Remakes Number	Last Year Remake
1	Casablanca	Weir	1942	3	1940
2	Dead Poets Society	Curtiz	1989	0	NULL
3	Rman Holiday	Wylder	1953	0	NULL
4	Sabrina	NULL	1964	0	1985

typing error

incomplete

outdated

inconsistent

Introduction & Motivation

- According to Dey [3]: *"**Context** is a general term used to capture any information that can be used to characterize the situations of an entity."*

- In particular, some DQ dimensions (accuracy, completeness, consistency, etc.) are characterized by (or depend on) different aspects:
  - type of application
  - users requirements
  - the task

- In the literature, there is no agreement about the degree of dependency between DQ dimensions and these aspects.

Introduction & Motivation

- We are inspired by the **Data Quality Management Process** of **AGESIC**, applied in the **Digital Government** domain.

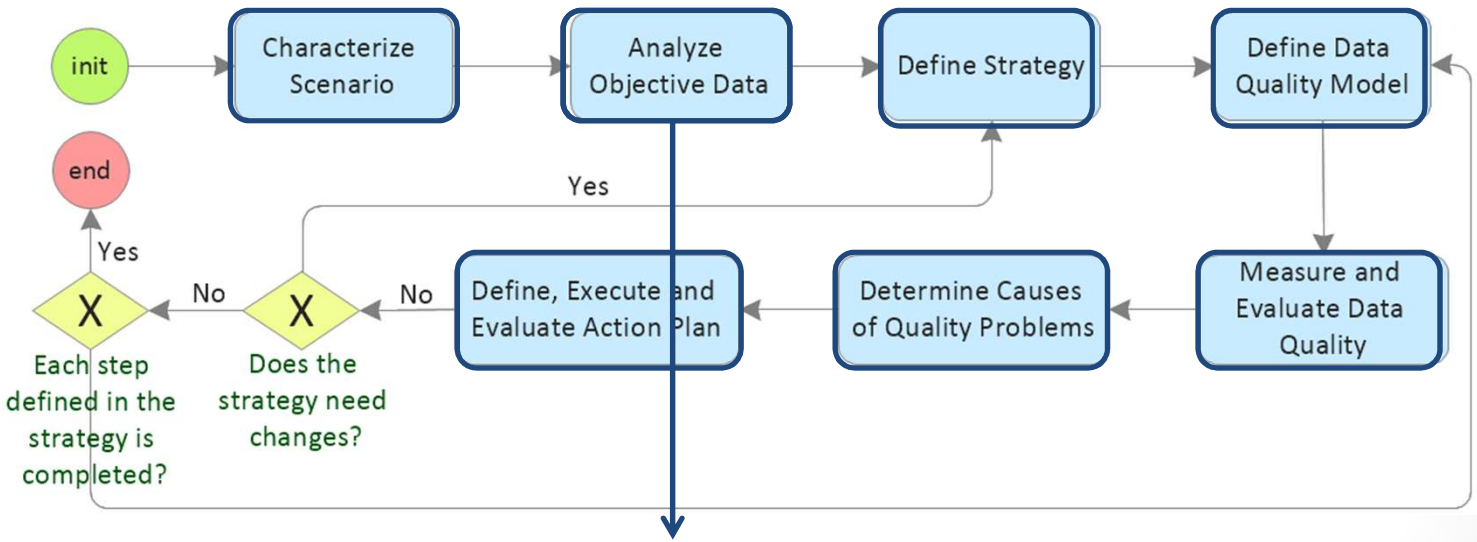
- Digital Government involves several actors: organizations, business processes, public services and citizens.

- AGESIC is the *e-Government Agency and Information and Knowledge Society* in Uruguay [4]. It is establishing DQ process standards to be applied by public bodies.

Introduction & Motivation

- DQM Process in Digital Government for AGESIC.

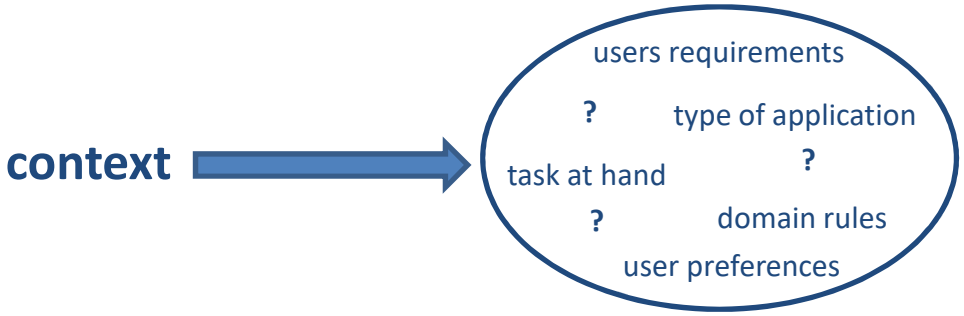
DQM Stages are defined by a set of Tasks



Task example: Data profiling technique (analyze number of null data, number of duplicates, etc.)
Some DQ problems arise at this stage.

Introduction & Motivation

- Our **objective** is to model context for each DQM process stage.



- Firstly, a review of the **State of the Art** was necessary.

SYSTEMATIC LITERATURE REVIEW

Systematic Literature Review

- **It is a methodology** to search bibliography.

- A Systematic Literature Review (SLR):
  - defines research questions to determine criteria for selecting relevant data to answer such questions.
  - provides a high level summary of the literature in fields connected.

- The scientific works found with a SLR are called **primary studies** (PS).

Systematic Literature Review

- Objective: **Relate** the following areas **Data Quality** and **Context**.

- **Research Questions:** Which works deal with…
  - **RQ1:** context and data quality models?
  - **RQ2:** context and quality metrics for the main data quality dimensions?
  - **RQ3:** context and data quality concepts?

- quality dimensions categories defined by Wang&Strong (1996) [1]
- the quality dimensions of the ISO / IEC 25012 Standard [6]
- the most important data quality concepts

9 search strings to execute in the Digital Libraries

SLR: Executed Process

To execute Search Strings in Digital Libraries

(Context OR preference OR "data tailoring") AND ("data quality model")

apply inclusion criteria:
- published since 2010 inclusive
- in English language
- must be an article or book chapter
- in pdf format

ACM	IEEE	Science Direct	Springer	Google Scholar
257	5	934	884	52

returned: 2132 primary studies

remove duplicates

returned: 1797 primary studies

select by relevance (title and abstract)

returned: 246 primary studies

select by full text

43 primary studies

apply exclusion criteria:
- abstract in English, full text in another language
- addresses quality, but not data quality
- data quality is very superficially
- does not address data context

review completed January 2020

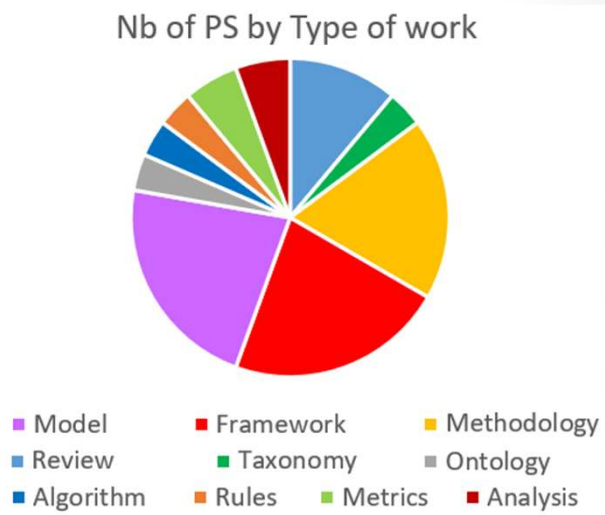
SLR: Data Analysis

- Analysis axes:
  - **Type of work**: review, taxonomy, framework, methodology, etc.
  - **Research domain**: Big data, e-Government, Internet of Thing, general, etc.
  - **Context definition**: formal, not formal, none
  - **Case study**: real data, non-real data, none
  - **Case study data model**: relational, graph, olap, etc.
  - **Restriction to data types/model**: structured, semi-structured, attribute values, cost value, general, etc.
  - **Venue quality**: in accordance with rankings and metrics of the Scopus journal
    - https://www.scopus.com
    - http://portal.core.edu.au/conf-ranks/

SLR: Some Results

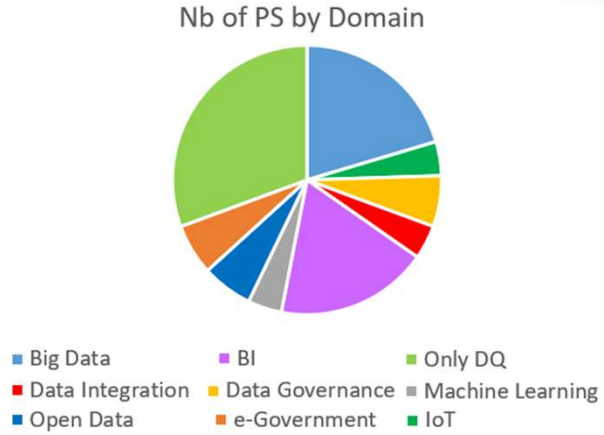
- Almost half of the selected PS were returned with the Search String that relates context and data quality concepts.
  - Most of them comes from Springer.

- Concerning types of works, most PS, propose:
  - Models,
  - Frameworks,
  - Methodologies for DQM



SLR: Some Results

- Most PS are in the following areas:
  - Big data,
  - Business intelligence,
  - DQ in a general way.



- Interestingly, the number of published papers dealing with the **use of context for DQ increased from 2016**.

- An important result is the **lack of works formalizing context**.
  In 43 PS selected, only 5 works propose formal definitions of context.

THE PHD PROJECT

The PhD Project

- Based on the SLR results:
  - most research **does not define** what context is.
  - in general, researches present an **informal** context definition.
  - there are very few researches that **formally** define the context used.

- We draw our **first research problem**: **Which components should be included in the definition of context for DQM?**

The PhD Project

- According to the SLR, the context could be defined by the following elements:

Users: Profile, preferences, task

Domain rules. For example:
$\forall t \in R$: (t.lastYearRemake > t.year)

DQ requirements. For example:
Currency(e-mail) < 30 days

DQ metrics. For example:

$$\text{Completeness} = 1 - \frac{\text{Number of tuples in R with age in NULL and e-mail in NULL}}{\text{Total number of tuples in R}}$$

The PhD Project

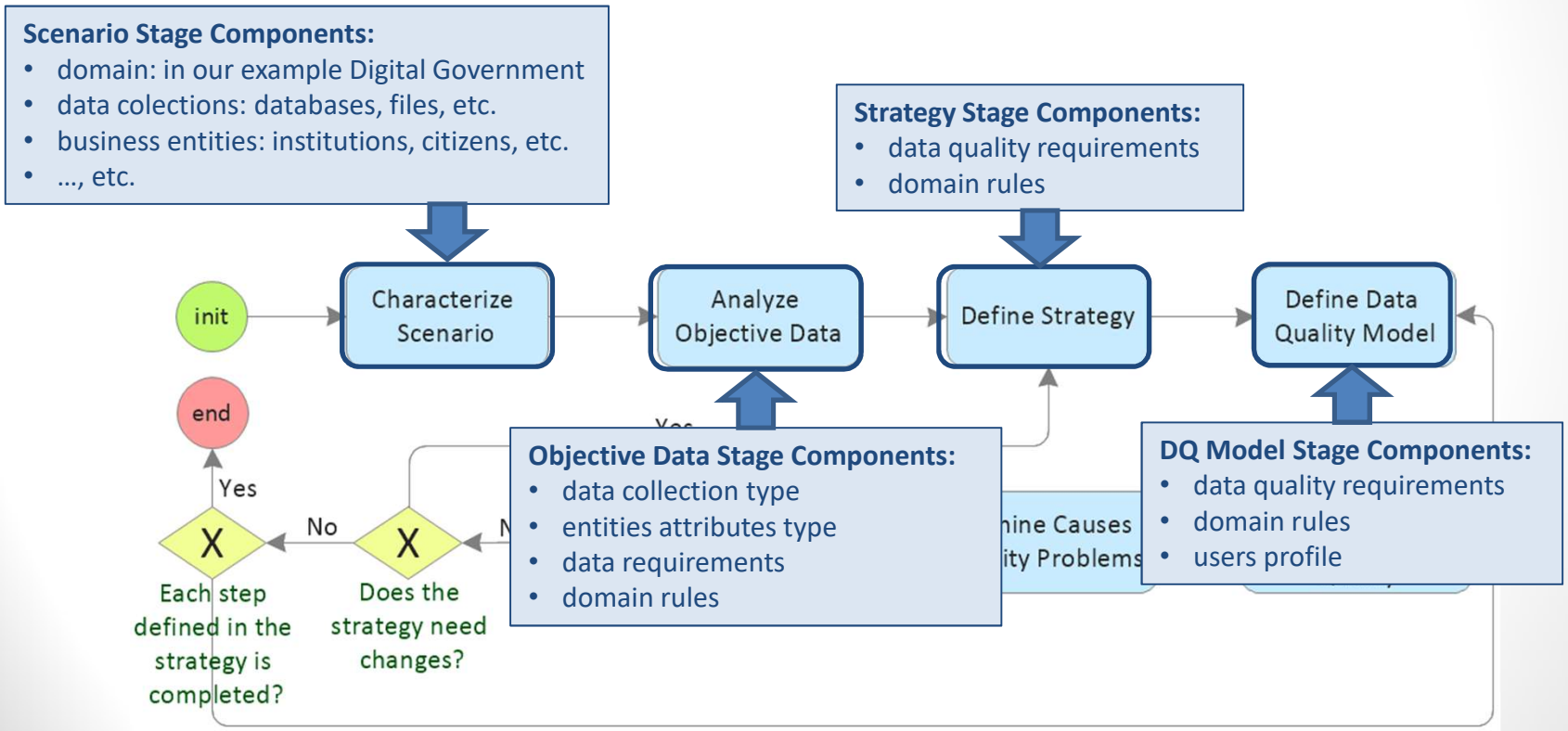
- We draw our **second research problem**: **How context components should be included in each DQM process stage?**

- This implies:
  - define a context for each process stage, determining all the components included in each context.
  - for each execution of the process, in each stage, instantiate the context.

- Define a **Formal Model of Context** for DQM process.

The PhD Project

- Example: **Possible context components** in the DQM Process stages.



The PhD Project

- A **Case Study** will be developed within AGESIC environment.
- The **Context Model** will be applied in the DQM Process of Digital Government.
- **Instantiating our Context Definition** in each stage of this process.

Planning

- This thesis started in September 2019.

TASK	2019	2020			2021			2022	
	9-12	1-4	5-8	9-12	1-4	5-8	9-12	1-4	5-8
SLR									
define context components									
model context for DQ process stages									
evaluation protocol									
experimentation in a real case									
manuscript writing									

References

- [1] Wang, R.Y., Strong, D.M.: Beyond accuracy: What data quality means to data consumers. J. of Management Information Systems 12(4) (1996).
- [2] Batini C., Scannapieco M. (2016) Information Quality in Use. In: Data and Information Quality. Springer.
- [3] Dey, A.K.: Understanding and using context. PUC 5(1) (2001).
- [4] https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/ Last access July 2020.
- [5] Kitchenham, B.: Procedures for performing systematic reviews. Keele university.technical report tr/se-0401 (2004).
- [6] https://iso25000.com/index.php/en/iso-25000-standards/iso-25012 Last access July 2020.

Thank you

ANY QUESTION OR SUGGESTION?