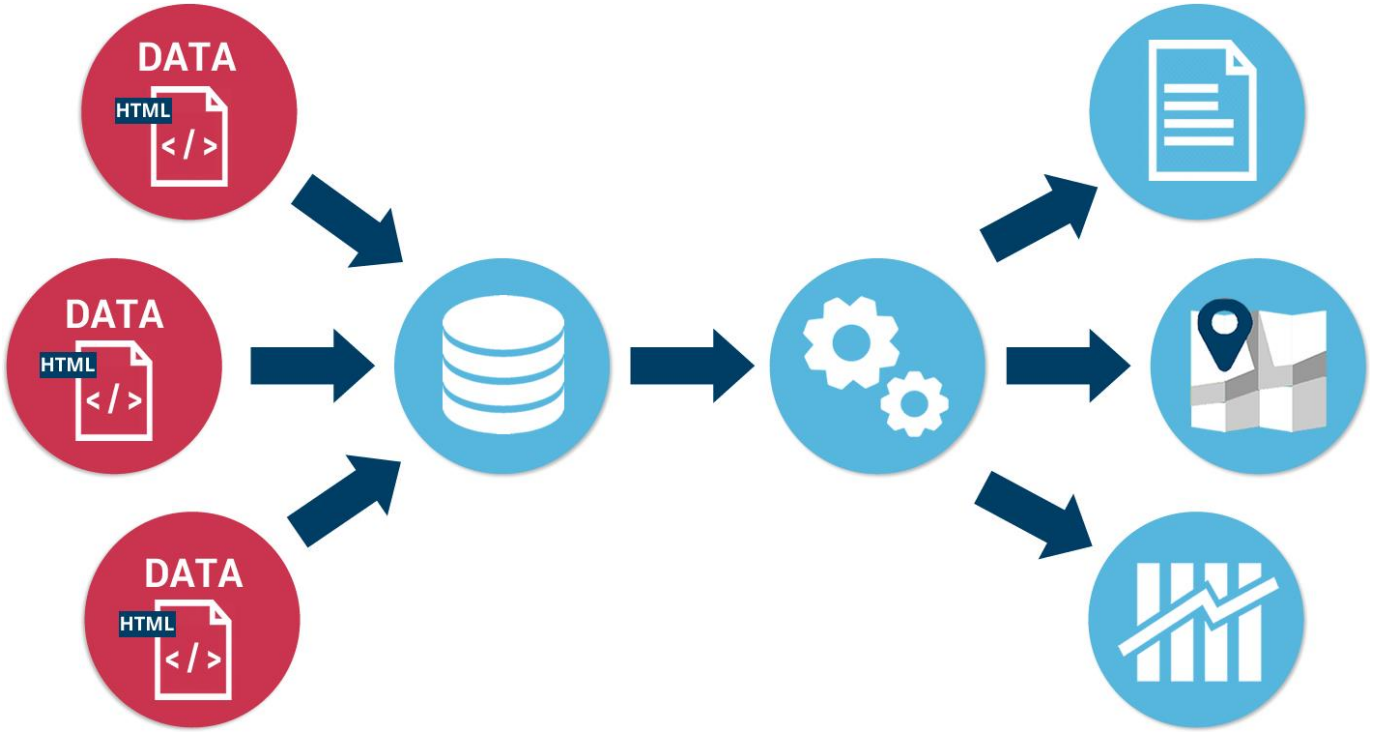
News bot for the **newsroom**

How building data quality indicators can support journalistic projects relying on real-time open data



Building quality indicators for automated data-driven projects

Laurence Dierickx – Université Libre de Bruxelles - 2017

Theoretical **backdrop**

“QUALITY DATA LEADS TO QUALITY INFORMATION”

- Data quality difficult to define
- Data quality literacy: multidimensional concept
  - Formal characteristics
  - Empirical data evolve with time
  - Journalism: reliability, objectivity, accuracy, believability, trustworthiness…
- ISO 9000: ability to satisfy implicit or explicit needs

FITNESS FOR USE PRINCIPLE

Building **data quality** indicators

FOR A DATA-DRIVEN JOURNALISTIC PURPOSE

Application domain: automating data
(scraping and/or analysis project – bots to support investigative)

Three stages assessment:

- The technical challenge (Boolean)
- The journalistic challenge (Boolean)
- Global assessment model (empirical)

TAYLOR-MADE TO FIT THE NEEDS OF COMPUTATIONAL JOURNALISM

The **technical** challenge

FOR AN OPTIMAL REUSE / FREE OF BIAS

To fit the development requirement

Four axes:

- Documentary (understandability of data-sets)
  - Encoding (technical encoding of data)
    - Normative (standards, norms)
- Semiotic (syntactic and semantic coherence of data)

code	gemeente	17/10	18/10	19/10	20/10	21/10	22/10	23/10
41B004	Brussel (Sint-Katelijne)	31	44	58	51	51	49	37
41B006	Brussel (EU Parlement)	NA	NA	NA	NA	NA	NA	NA
41B011	Sint-Agatha-Berchem	45	65	71	65	70	66	57
41N043	Voorhaven (Haren)	NA	22	54	55	59	58	34
41R001	Sint-Jans-Molenbeek	39	57	63	61	62	58	42
41R012	Ukkel	44	58	65	60	62	55	44
41WOL1	Sint-Lambrechts-Woluwe	36	47	55	58	57	56	36
42N016	Dessel	NA	NA	NA	55	NA	39	47
42N027	Bree	60	74	72	68	71	65	60
42N035	Aarschot	45	73	68	66	70	64	48
42N040	Sint-Pieters-Leeuw	57	61	71	61	68	64	59
42N045	Hasselt	NA	NA	NA	NA	NA	NA	NA
42N046	Gellik	47	74	65	73	66	67	54
42N054	Walshoutem	41	63	69	67	63	59	51
42R801	Borgerhout	NA	NA	NA	41	54	52	30

Formal DQ **indicators**

Axis	Assessment types
Documentary	Unique identifier Available metadata Conformity metadata/data-set Terms of use
Encoding	No encoding problem No HTML overload No duplicate data
Normative	Use of standards (e.g. e-mail, date, address, geolocation,…)
Semiotic	No missing value No orthographical incoherence Explicit labelling

The **NULL** value

DIFFERENT POSSIBLE INTERPRETATIONS

- Information exists but is not known
- Information is not relevant for the entity
- Information is relevant but does not exist for the entity
- Attribute value is equal to zero
- Poor quality data can coexist with correct data without generating errors
- Source may contain any error but that the data does not have the meaning expected by the user.

The **journalistic** challenge

DATA QUALITY DIMENSIONS AND EMPIRICAL INDICATORS

INTRINSIC & CONTEXTUAL DIMENSIONS

Meet the journalistic preoccupations about accuracy, currentness, precision and completeness.

! WARNINGS !

“Completeness” indicator might be more difficult to assess due to the issue of the NULL value

A primary source (refers to the responsible for the control and of the monitoring over time of the quality of the information produced)
does not guarantee the quality of a data-set but only give indications about it

DQ dimensions **indicators**

Dimension	Assessment types
Contextual	Primary Source (authentic)
	Appropriate amount of data
	Completeness (no missing values)
	Relevance
Intrinsic	Accuracy (syntactic correctness)
	Precision (no anomalies observed in values)
	Correctness (f.e, last update mentioned)

Global **assessment** model

Empirical indicators cannot be avoided

= to remain critical

Four axes:

- Reliability of the data source
- Modalities to access to data
- Keys to understand data and their production context
- Automation allowed
- Relevance for journalistic use

Empirical **indicators**

Axis	Question to answer
Source	Is the data provider the producer and/or the authentic source? In the case of the data provider is not the original producer and/or the authentic source, what is the nature of its relationship with the original producer of the data and/or the authentic source? Are the data provider, the data producer and the authentic source of data trustworthy?
Access	Are data freely accessible? Are they licensed for free reuse? Are they available in a structured format?
Automation	Are data provided in a free and usable format? Do the data values meet the standards? Are the data values accurate? Is the data-set complete and up-to-date?

Empirical **indicators**

Axis	Question to answer
Documentation	Are data documented by metadata or any other type of information which permit to understand the structure of the database and/or to remove any ambiguities in the data labeling? Is any expertise provided to understand what data values are? Are contextual elements provided?
Journalistic relevance	Do data have an added value in journalistic terms? How does the data processing make sense?

DATA QUALITY INDICATORS COULD VARY FROM ONE DATA –SET TO ANOTHER

Case study: **Bxl'air bot**

Domain: air quality in Brussels (PM10, PM23, BC, O3, NO2)

Object: automation of the collect and the analysis of data as raw material for a wider investigative

LESSONS LEARNED…

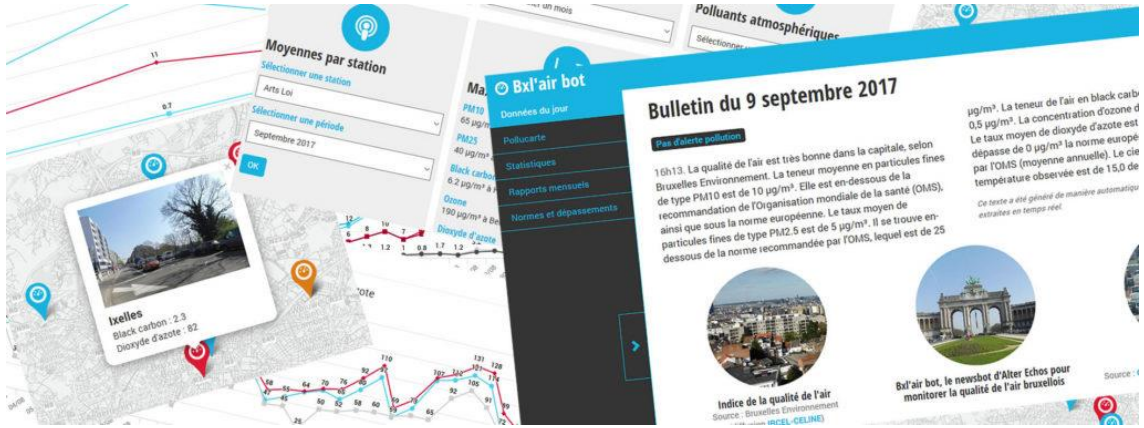
- Highlight strengths and weakness of a data-set
- Right technical decisions made, preventing errors
- Permitted to make the most relevant choices for the purpose

BUT…

Case study: **Bxl'air bot**

- Does not explain how data evolve with time
- Necessity to have a deep knowledge of the application domain

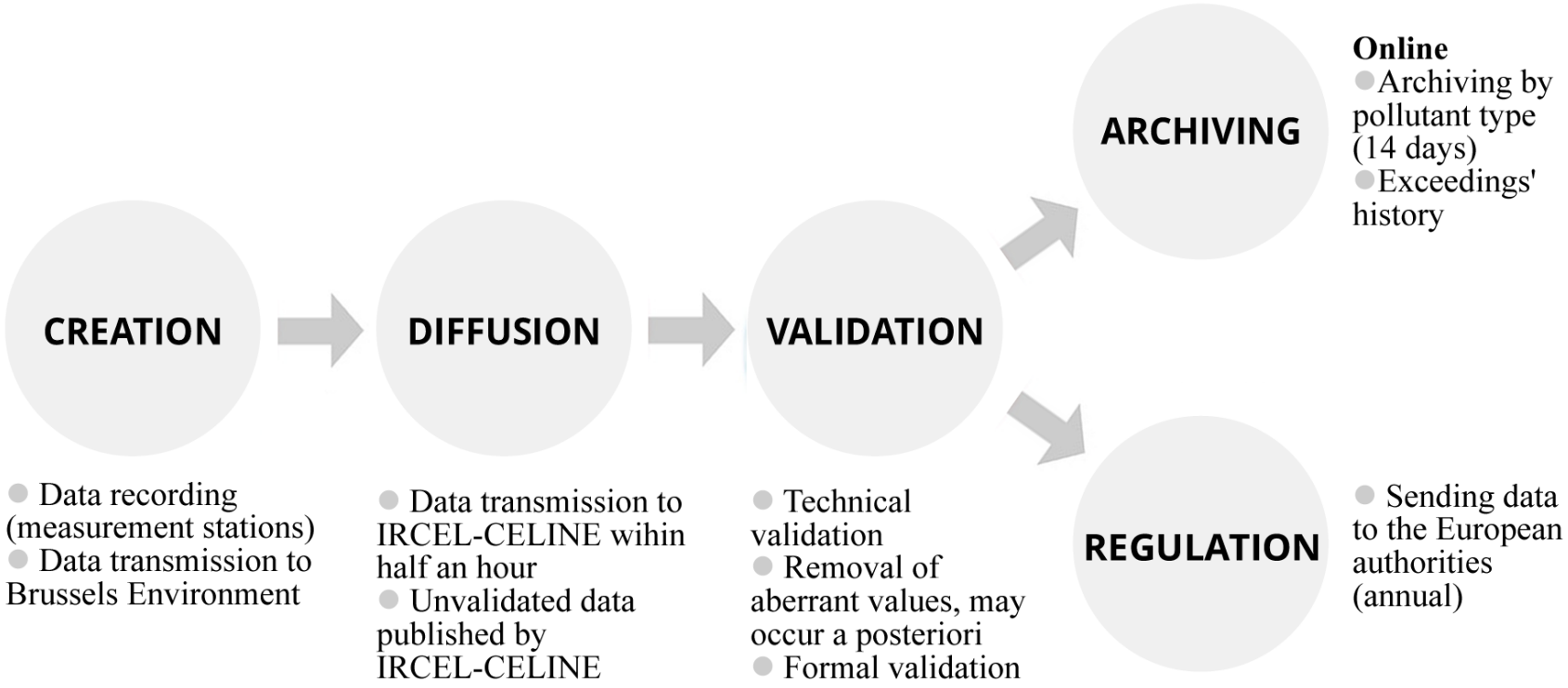
ASK THE EXPERTS, THEY ARE VALUABLE!



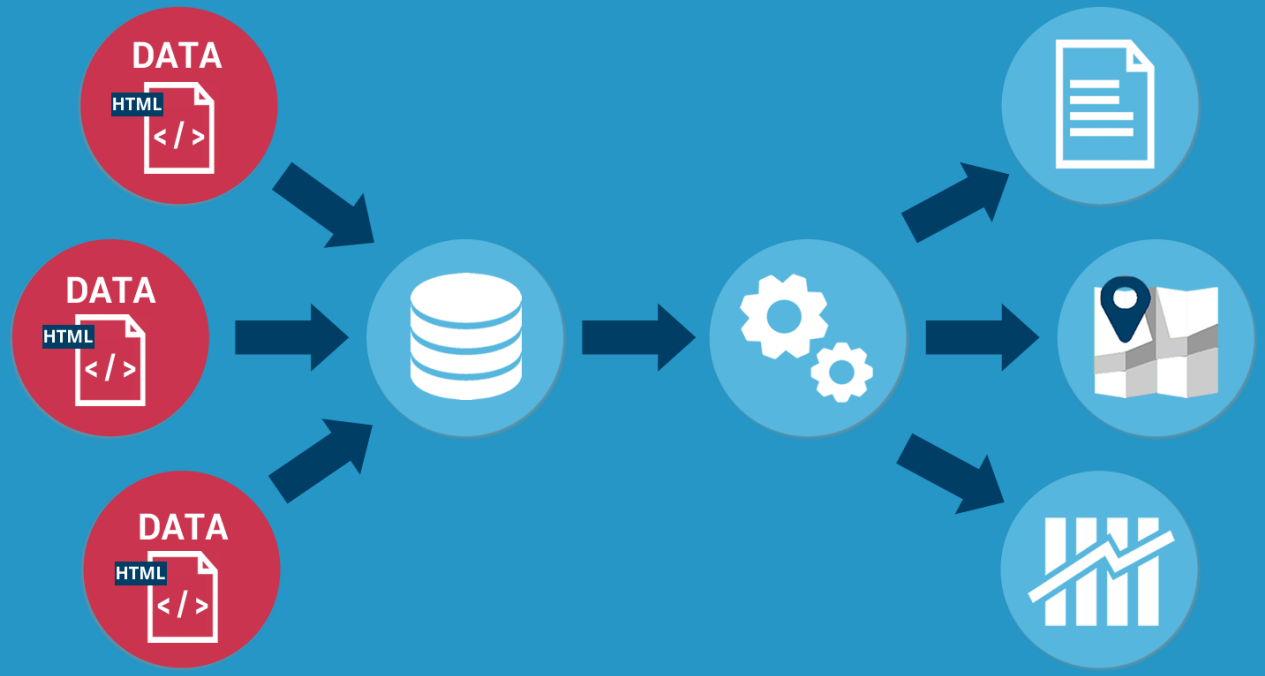
www.bxlairbot.be

Data **Life Cycle**

How are data building with time?



Thank you!



www.ohmybox.info

@ohmyshambles