Pôle Systèmes d'information de

SIG | Data | Si collaboratif | SI décisionnels - IA | Supervision

Generic evaluation of the arc-node data type quality applied to water and waste-water networks

Christian CAROLIN, Mathieu LE MOAL
(AXES Conseil)

4th International Workshop on Spatial Data Quality
Bruxelles · 11-12/10/2023

Presentation

- About AXES Conseil
- Why data quality is a major issue?
- Process
- Indicators
- Next steps

AXES Conseil, Information system unit of **SEGIC Ingénierie** :

SEGIC Ingénierie is major engineering player in the fields of Environment and infrastructures in France.

Areas of intervention

>Urban planning
>Road
>Civil Engineering
>Landscape
>Transport
>Water and waste water
>Environment
>Energy; equipment
>**Information systems**

AXES Conseil

	**GIS**	Project Management and assistance for GIS project implementation : specification, strategic study, software choosing, testing, etc.
	**Data**	GIS data quality evaluation and GIS data processing
	**Collaborative**	Advise and assist in setting up collaborative tools and standard or personalized content management systems
	**Decision-making and AI**	Methodological and/or technical assistance for the choice and integration of geo-decision systems and decision-making information systems, including AI
	**SCADA**	Methodological and/or technical assistance for the choice and integration of supervising system (SCADA projects)

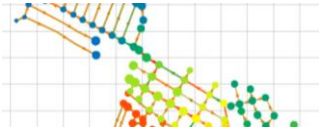
Why data Quality is a major issue ?

Growing environmental challenges make **water management a crucial subject** for our society.

The **quality of databases** describing water or wastewater networks becomes **a major issue.**

Water and wastewater operators need **reliable data to make the appropriate decisions :**

- Hydraulic Modeling
- Works planification
- Locating network weaknesses
- Etc.



Why data Quality is a major issue?

There is no “absolute” data quality. **Quality is directly linked to usages** (and so, the users). So, let users assess their data make sense !

Data quality assessment could be realized by users : data must match with their uses and aims.

But users usually :
> don’t know (or partially know) the data specifications
> are not still able to use specific data quality software
> lack time

With some **tangible indicators, users can analyse the data according to their practice**. Examples :

- Choose available values for the diameter,
- Define the minimal attribute fill rate acceptable,
- Define the minimal distance between two pipe manholes,
- Select the minimal distance between two pipe manholes

Why data Quality is a major issue?

Current situation about GIS data quality, according to our experience

> Databases are often old, with outdated structures

> Updates are heterogeneous, rarely based on a formalized process

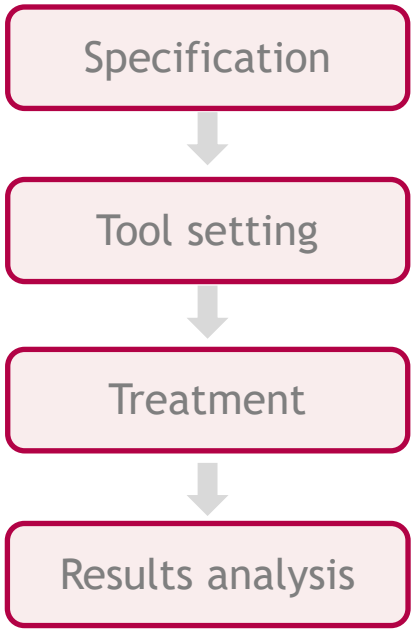
> Data quality management software are often expensive, complex and time-consuming for users

Based on these observations, AXES Conseil has developed a new approach.

Process

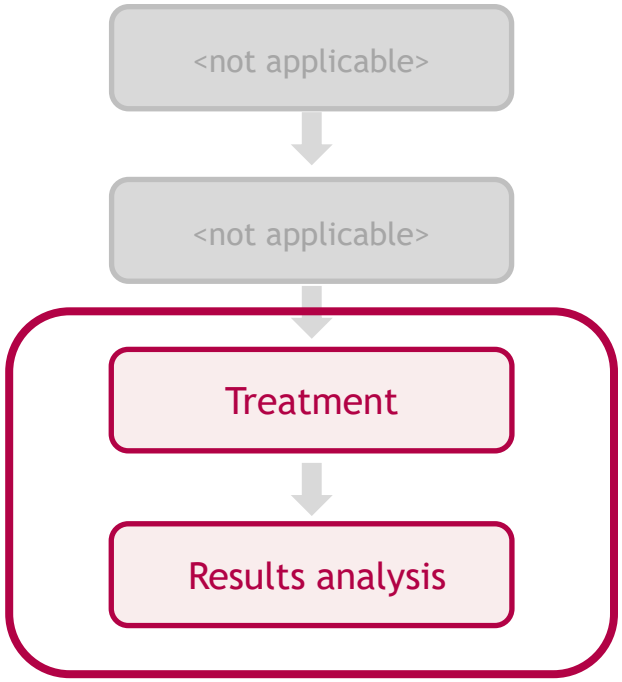
Current approach

- Current quality control solutions proceed by translating specifications into computer rules on dedicated tools.



AXES Conseil's approach

- AXES Conseil's approach is to generate and then analyze a **set of generic data quality indicators**



Process

Process duration : ~ **2 to 4 days**

1 : Data are provided (files format)

- Set the parameters (select the attributs, file format, etc.) with a short meeting
- Available input formats :

SQLite GEOJSON .SHP

2 : The data is analyzed with PostGRES/PostGIS's dedicated scripts

- The **analysis is exhaustive**: all objects are processed

PostGIS Spatial PostgreSQL

3 : The results are delivered (report, datasheets…)

- Outputs :
  - SQLITE file : data & indicators,
  - QGIS project : visualisation (map),
  - Excelsheet : synthesis.

Possibility of periodic comparative analysis

Process

Benefits of this new approach :

- **No expensive tool(s) needed** (data quality software, ETL…). It's a service.

- Can be performed **without knowing data specifications.**

- **Not training required** on the tool. User use his own GIS tool to manipulate output data

- **Fast process**

- High level **flexibility** during the analysis step

Indicators

Quality indicators are organized into 6 categories :

MCD	**Modelisation**
STR	**Structure**
GEO	**Geography**
REL	**Relation**
ATT	**Attribut**
RES	**Network**

For each category, there are many indicators.

Each indicator is illustrated below.

Indicators

From **~15 to 25** indicators are generated (according to selected indicators).

For **1 input table**, from **~8 to 15 output tables** are populated (according to selected indicators).

The indicators partially cover those of **ISO 19158** standard relating to data quality.

Current quality indicators typology	
Genealogy	No
Temporality	No
Geography (location)	Partial
Geometry	Yes
Completeness	Yes
Semantics	Yes
Consistency	Yes

Indicators illustration

1- Modeling [MCD]

Indicators related to the structure of tables and attribute values *

- Table and attributs name structure
- Values structure
- Geometry type

(*) : only avaible for the SQLITE format data input

Value structure

>Is space character present ?
>Is special character present ?
>Etc.

le nombre de valeurs uniques est :	5	La valeur SENS dispose de 5 valeurs uniques
le nombre de caractères min est :	6	6 caractères au mini
le nombre de caractères max est :	7	7 caractères au maxi
Les valeurs sont homogènes pour tous les objets :	1	valeurs homogène (même type)
existe-t-il des caractères spéciaux (autres que a9)	0	Pas de caractères spéciaux
Existe-t-il des caractères 'ESPACE' en début ou fin de valeur	0	Pas d'espace en bordure de valeur
Existe-t-il des caractères 'ESPACE' dans la valeur	0	Pas d'espace dans la valeur
La valeur est elle homogène a9 ?	0	
Nombre de valeurs NULL	193	
Ratio de valeurs NULL	9,2	

Table and attribut name conformity with PostGRESQL recommandations

Nom de l'attribut non conforme	Type de l'attribut
alti_GS	real(1000,500)
125hz	double
AT_100	double
Liste(a)	varchar(255)
emetteur_de_la_DT	varchar(255)
References_Cadastrales	varchar(128)

Geometry type

nom_base	*item*	*liste_valeurs*	*nb_valeurs*
Base1	coord_dimension	2	169
Base 1	coord_dimension	3	5
Base1	f_geometry_column	geom	174
Base1	geometry_type	multipoint	97
Base1	geometry_type	multilinestring	33
Base1	geometry_type	multipolygon	39
Base1	geometry_type	z-m	1
Base1	geometry_type	z-m	3
Base1	geometry_type	z-m	1
Base1	srid	3943	173
Base1	srid	4326	1

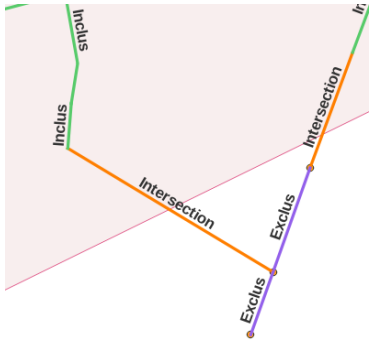
Indicators illustration

2- Geography [GEO]

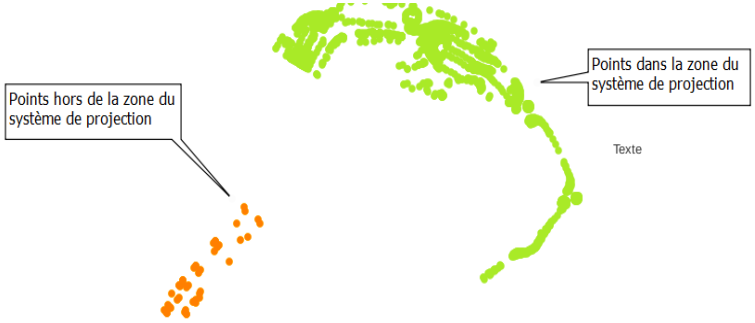
Projection and coordinates (x,y) analysis

- Data coordinates digit
- Compliance Compliance of the position of objects on a given footprint
- Projection range

Compliance of the position of objects on a given footprint (ex : town area)



Compliance with the coordinate projection range



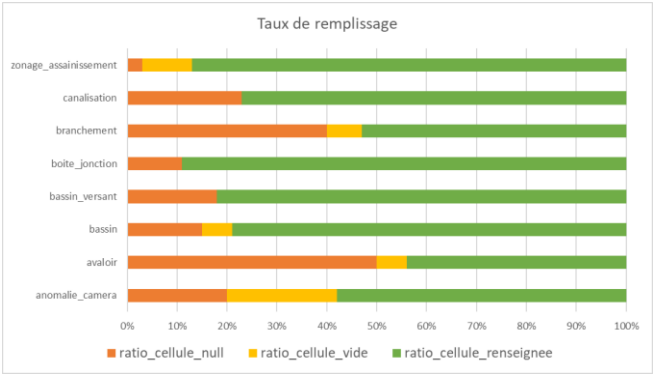
Indicators illustration

3- Attribut [ATT]

Attributs and values indicators

- Fill rate (table, attributs) and Fill rate location
- Confusion matrix
- Values statistics
- Etc.

Table or attribute fill rate

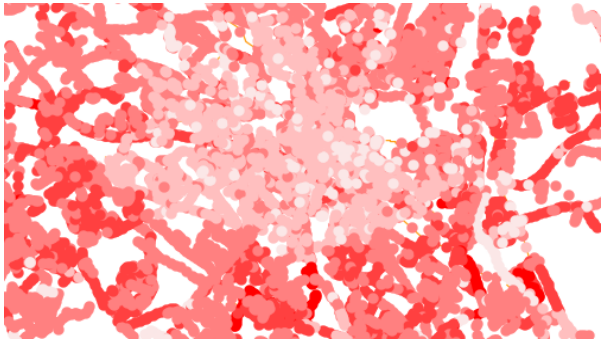


Confusion matrix (2 values)

Ex : date / material

nb tronçons	Matériaux		
Annee	ACIER	FONTE	PVC
1967	10	25	65
1968	14	16	1
1969	15	60	
1970			22
1971			17
1972			4
1973			6
1974		1	2

Attribute Fill rate location



Indicators illustration

4 – Structure [STR]

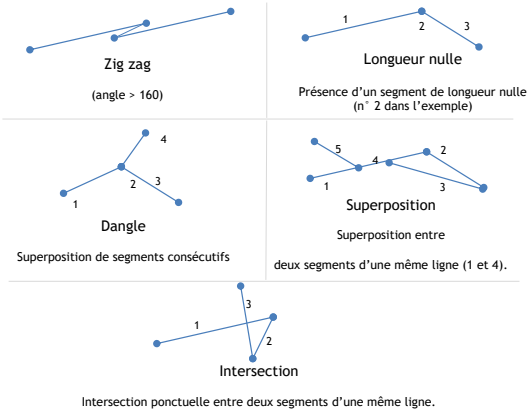
Object geometry analysis

- OGC conformity
- Geomety singularity (angle, null length, etc.)

>Each error / singularity is located

OGC conformity

Geometric singularities



Indicators illustration

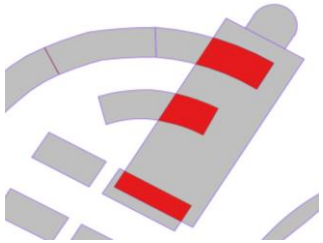
4 – Spatial relationship [REL]

Spatial relationship between objects from the same layer

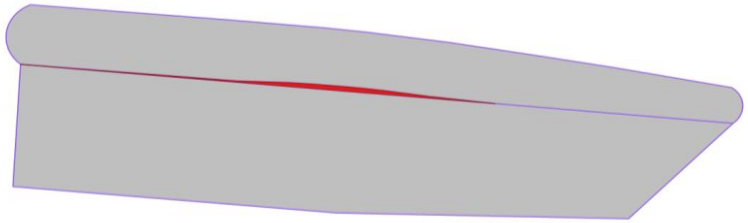
- Line intersection
- Interstices and overlapping polygons

>Each « relationshionship « is located

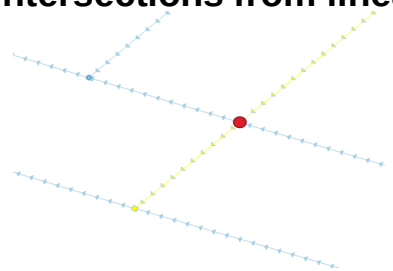
Overlapping



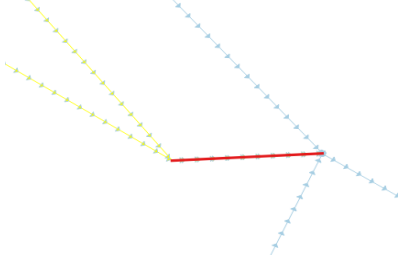
Interstices



Ponctual intersections from linear objects



Linear intersections from linear objects



Indicators illustration

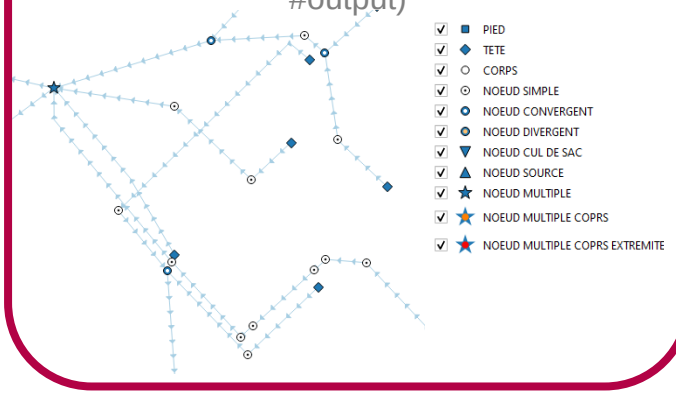
5 - Network [RES]

Arc Node data analysis (oriented or not)

- Node typology depending on connection configuration
- Potential discontinuity
- Attribut values variations at nodes
- Topological unit
- Etc.

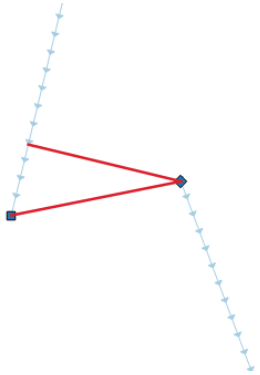
Typology of nodes and elements

(based on connection configuration #input / #output)

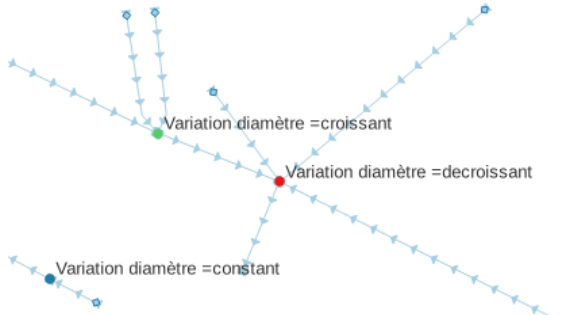


Topological unit

Potential discontinuity



Attribut value variation at each connexion (ex : diameter)

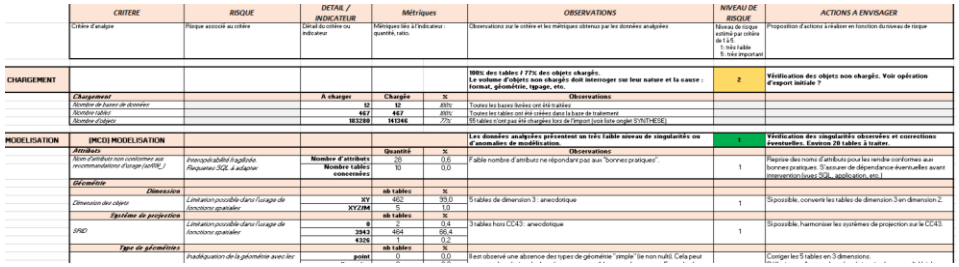


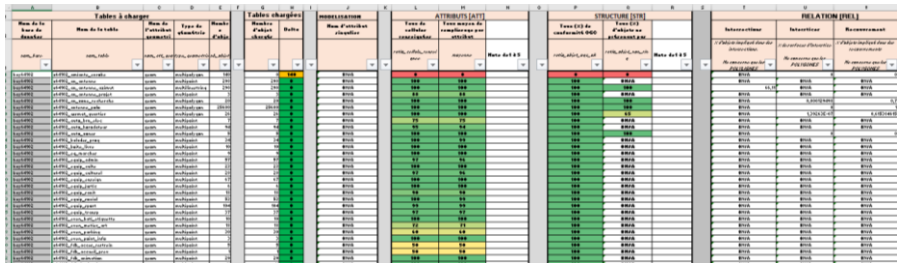
Indicators illustration

The treatment outputs are :

- SQLITE file containing all tables (data source + indicators)
- QGIS project(s) to map results

- Excel sheets summarizing results





Observations

The following observations are based on more than 20 references achieved.

According the massive data produced, **presentation method is essential** to allow users to understand and then define the way to analyse data.

- This topic is one of the items addressed by the **French national Data Quality Working Group : QUADOGEO** (https://cnig.gouv.fr/gt-quadogeo-a18183.html).

Indicators analysis requires **real users involvement** to be completely efficient.

The variety of indicators calls different type of answers :

- Short term
  - Ex : fix the geometry error (dangle, etc.)
- Long term
  - Ex : Modeling, modification

Following analysis, a real **“Data enhancement” project has to be defined** : actions, plan et resources. Data quality evaluation is the start point of real entire project.

Next steps

Improve results presentation

> It’s the key to allow users to go in the subject.

Tracking data improvement

> As uses change, spatial analysis become essential for decision-making

Explore the contribution of the indicators in prediction (AI)

> “Right equipment at the right place” or “equipment could be missing here”

The presentation focused on the water and wastewater theme, but the concepts presented are equally applicable and applied to all "Arc-Node" type models, such as gas, electricity, roads, etc.

Thank you for your attention.

Contact us

AXES Conseil Pôle Système d'information de Ségic Ingénierie

7 rue des Petits Ruisseaux
F-91370 VERRIERES LE BUISSON

http://www.axes.fr / axesconseil@axes.fr

Mr. Christian CAROLIN : carolin@axes.fr

Mr. Mathieu LE MOAL : lemoal@axes.fr