Towards a generic fusion framework for underground networks involving model-driven engineering domain

Yassine Belghaddar[1,2,3,4]
yassine.belghaddar@berger-levrault.com

Abderrahmane Seriai[1]
abderrahmane.seriai@berger-levrault.com

Ahlame Begdouri[3]
ahlame.begdouri@usmba.ac.ma

Carole Delenne[2,4]
carole.delenne@umontpellier.fr

Nanee Chahinian[2]
nanee.chahinian@ird.fr

Bachar Rima[1]
Bachar.Rima@lirmm.fr

Mustapha Derras[1]
mustapha.derras@berger-levrault.com

[1]*Berger-Levrault*, Pérols, France
[2]*HSM, Univ. Montpellier, CNRS, IRD*, Montpellier, France
[3]*LSIA , Univ. Sidi Mohamed Ben Abdellah*, Fez, Morocco
[4]*Inria Lemon, CRISAM – Inria*, Sophia Antipolis Méditerranée, France

***Abstract*—Underground networks, particularly sewerage networks require accurate information for their management. To successfully and smoothly accomplish the required tasks for network expansion, reparation or simulation analysis, operators collect and analyse data coming from multiple sources. The various and heterogeneous sources available often provide different representations for a network. Thus, raising challenges for information exploitation and communication between the different actors managing these networks. In addition, the imperfections related to the sources are numerous and their consideration in the decision making process is mandatory. In this paper, we propose a generic data modelling for the fusion of sewerage networks data. Our meta-model supports imperfection modelling at data-source level as well as at network object position and attribute levels, allowing thus formal fusion operations to be conducted efficiently and reliably. To validate our meta-model, we implemented it using data analysis and reengineering platform called Moose, and we conducted a test on the town of Prades-le-Lez (France). We took into account three data-sources providing information on the node positions of the sewerage network : 1- the official network map as semi-structured source, 2- a high resolution aerial image database and 3- a Google Street View database as unstructured sources. As result, we were able to reliably perform data monitoring and visualization requests on real heterogeneous multi-source data related to a specific sewerage network.**

***Index Terms*—Meta-modeling, Sewerage Network, Model-driven engineering , Data sources**

I. Introduction

By the time a sewerage network is set up, its graph (where the nodes represent the manholes or inlet grates and the edges represent the pipes) is mapped for the first time. Later on, several actions, such as repairs or expansions, may occur in the field according to the new needs of the citizens [1], [2] . These modifications to the network are recorded by the operators that have conducted the actions. The updated data on the sewerage network are generally stored in several formats and representations such as images, text documents describing the activities, SIG/shapefiles, csv files, etc. For example, data on the sewerage networks published on the French Government's open access portal [3] shows this diversity. Consequently, the combination of data from different sources and eras raises problems of consistency (data conflict), which may be due to differences in granularity or accuracy of the data sources [4] and requires the establishment of a methodological framework for collecting, centralizing, updating and data archiving in order to facilitate information sharing and communication between the managers. In our vision of data fusion, we consider the uncertainties related to each type of collected information in order, for example, to anticipate and react promptly to potential dysfunctions or to quantify their impact on the results of a numerical simulation of flows in the sewerage network. In this context, and as a first step of our work, we propose in this paper a meta-model for aggregation, control and analysis of data sources related to sewerage networks, before elaborating adapted algorithms to merge heterogeneous multi-sources data. This paper is an extension of the work described in [5], it is structured as follows, Section II introduces the context of our work. We explain the motivation behind this work in Section III and we present the state of the art and related works in Section IV. Section V describes our meta-model and its specific viewpoints. Section VI demonstrate the implementation and the results of instantiation of our meta-model in Moose [6], using real data of a sewerage network provided by multiple sources. Section VII concludes this paper.

II. Context

A. Sewerage networks and management challenges

Sewerage system is a network for collecting and transporting wastewater and storm water to a treatment plant, also called combined sewer system. When a network collects these two types separately, it is called separated sewer system. To set up a sewerage system and make its progress in a territory, different institutional and operational actors

are involved. The ministry in charge of sanitation or the ministry in charge of local communities, define sanitation policies and strategies as well as the regulatory framework at national level. The local authorities ensure the respect of regulations related to the quality of sanitation services. Finally, the contracting authorities (municipalities or state agencies) are responsible for the development of services, their quality and sustainability. For the implementation, monitoring and control of these services, they call for actors such as service operators, local associations, and development partners (founders, NGOs, and design offices). The sanitation supply is not only limited to infrastructure installation. Maintenance tasks such as reparation, expansion, damage anticipation and scheduling of interventions are all necessary actions to ensure a permanent and transparent services. Planning is an important task for decision making. It allows to develop a vision of needs in space and time, to quantify and prioritize them in order, among others, to direct funding towards the most necessary investments and at reasonable costs.

On a territory, infrastructures are large and must be managed in a collaborative way by the diverse involved actors. Urbanization and the concentration of populations in cities engender the increase of the dimensions of sewerage networks. For example, in France, this heritage consists of approximately 337,000 km of collectors [2]. The 2012 reform "DT-DICT" [7], as part of the network detection process indicates that France is covered by more than four million kilometres of networks, one third of which are aerial and two thirds are buried or underwater

The improvement of underground networks, particularly for sewerage networks, has several advantages, especially the impact on public health and environment preservation through the protection of water resources against pollution. The administrative management and the techniques of interventions play a key role in these challenges since they help to reduce damage costs induced from services interruptions and floods.

The expenses associated with the management of these infrastructures are high, particularly repairs, since the components of these networks are subject to degradation and damage caused by several factors: age, environment, etc. Furthermore, the costs of urgent and unexpected operations are far higher than the ones anticipated [2].

In this context, improving knowledge on the state of these networks, mostly underknown, becomes a priority. Indeed, digital technologies such as Geographic Information Systems (GIS) and Computerised maintenance management systems (CMMS) bring great added value and are officially increasingly adopted. For example, the regulations associated to the environment code in France require stakeholders to have digital and precise cartography for sensitive underground networks since January 1st, 2019 in urban units and from January 1st, 2026 in other cases.

These solutions are particularly useful for making spatial and geographic data available to the various actors, to facilitate their communication for optimal decision-making as well as to improve the administrative, economic, and financial management of this type of networks and of interventions to take place near these underground networks.

Since the majority of the components of sewerage networks are buried, collecting and detecting information about these infrastructures is challenging. However, with the advent of new technical solutions, different methods were used to extract information such as Ground Penetrating Radars [8]. Nevertheless, these propositions are usually applicable and efficient under certain conditions and constraints, such as the availability of Google Street View images in the case of [9]. Besides, each study usually focuses on limited components of the network. For instance, in [9], [10] only manhole positions are being collected. In [4], the focus is on collecting pipes positions. As for [11], the attributes of the objects are the target. Thus, a generic framework in which all these propositions could be integrated is necessary to help collect all the network data.

B. Sewerage network representation

A sewerage network is represented by a graph composed of nodes and edges. Nodes represent manholes, equipment, repairs, etc. the edges represent pipes. Each of the nodes and edges has a set of properties in the form of attributes such as, diameters of the pipes, types of materials and positions of the inspection areas for the objects. In recent years, storage and data management solutions for sewerage networks have evolved. Currently, most managers use Geographic Information Systems to create, edit, view, and analyze these data. The data structures and formats supported by these systems are diverse: relational databases, Shapefiles, GeoJSON and CSV files, etc. Moreover, to study the impact of some parameters, such as the discharge rate of consumers into the networks, specialists use hydraulic simulation software.

Although the applications are various, the digital representation of the data remains almost identical in the different solutions:

- Spatial data that are represented by geometric shapes and their relationships: points and lines (figure 1).
- Atributes that are listed in attribute tables where each record is associated with a network object (figure 2).

III. Problematic

When using GIS solutions, data processing includes acquisition, digitization, import, export, and visualization of geographical data. The acquisition can be carried out directly in the field allowing real time data collection and mapping. In addition, advanced data processing may allow considering multi-source data, spatial analysis through interactive queries and maps overlays.

Although the use of digital maps is increasingly adopted, there are still large communities in the world where maps and geographic data are still analog, making their use and update difficult. The detection and digital mapping of buried networks by semi-automatic or automatic approaches is a real scientific and technological challenge. Therefore, there is a large conceptual, technical and semantic gap between the analog and digital mapping models.

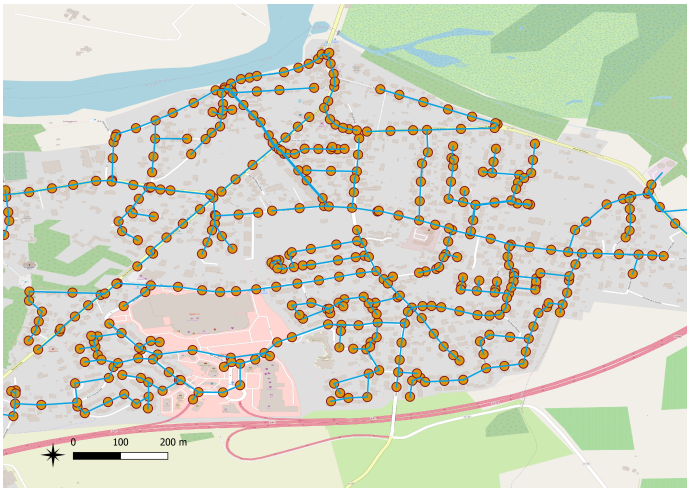

Fig. 1: Example of spatial representation of sewerage networks

	commune	ecoul	nom_voie	typer	dimensions	materiau	id
16	SAINT-JEAN-DE...	GRAVITAIRE	CAMILLE CLAU...	EAUX USEES	200	FONTE	15
17	GRABELS	GRAVITAIRE	NULL	EAUX USEES	999	INCONNU	16
18	MONTPELLIER	GRAVITAIRE	DU FAUBOURG ...	UNITAIRE	999	INCONNU	17
19	PRADES-LE-LEZ	GRAVITAIRE	NULL	EAUX USEES	200	INCONNU	18
20	LATTES	GRAVITAIRE	NULL	EAUX USEES	999	INCONNU	19
21	SAINT-JEAN-DE...	GRAVITAIRE	RUE CAMILL...	EAUX USEES	200	PVC	20
22	LE CRES	GRAVITAIRE	NULL	EAUX USEES	999	INCONNU	21
23	SAINT-JEAN-DE...	GRAVITAIRE	RUE CAMILL...	EAUX USEES	200	PVC	22

Fig. 2: Example of attributes table

The attributes and characteristics associated with the various objects constituting a network are not all available at a given time [11], [12] . This is partly explained by the fact that the networks undergo expansions and repairs but not properly tracked and reported, or through the interventions at different stages by actors, other than the operators who ensure the continuous functioning of the supply services. However, these attributes may be reported elsewhere, for example, in public databases, calls of tender, repair reports or even in press articles reporting damages.

In addition, since information and communications technology are easy to reach and use, operators currently have access to several sources from which they can collect useful data before interventions in the field, such as images, analogue maps, reports of interventions, sensors, etc. The heterogeneity of the sources makes the extraction of relevant information and its combination a complex and time-consuming task.

On the other hand, imperfections may be found in these data sets and sources, namely inconsistency (abandoned pipes which still appear on the maps), missing attribute values for some objects, uncertain and sometimes contradictory values. All of these aspects represent various obstacles to operators when merging the data.

Combining multi-source data also requires a unified data model to allow the centralization, updating, archiving, and monitoring of these data. Indeed, we have analyzed digital databases related to sewerage networks to understand the semantics of their data, their relationships and determine their differences. Since the associated data models, when they exist, are rarely available to the public, we proceeded by inferring them from data. In our study, we have used the data provided by reliable sources, particularly, the French open data repository [3]. Among the suppliers are the urban community of the South-East of Toulouse Sicoval, Data Angers, and the region of Pays de la Loire.

As a result of studying these different sources, we identified the following constraints:

- The data models adopted by operators are different. Therefore, exchanging and reusing data is difficult.
- The models do not comply with computer design and modelling rules and standards.
- The attributes provided by the stakeholders are related to their fields of activity. For example, a company specialized in hydraulic modelling provides precise information on the flow of water in a pipe, while this same information is generally missing in the data coming from another entity expert in the field of structures repairing.
- The history of interventions, necessary for anticipating repairs, is rarely considered in these models.

Thus, a generic model for business data and data sources is needed to overcome the conceptual and semantic gap between existing digital mapping models and to implement an optimal data fusion approach.

IV. State of art

A. Meta-models

Several definitions have been proposed. According to [13], a meta-model is a model used to model modelling itself. A Meta-model is a model that defines the structure of a modelling language [14]. The basic idea of a meta-model is to identify the general concepts in a given problem domain and the relations used to describe models [15]. This generality, which we also seek to satisfy in our proposal, is one of the most important axis that have made the meta-modelling, i.e. the creation of meta-models, one of the most important approaches for modelling. Instances of a meta-model are models that must satisfy the meta-model specifications. They enable target systems to be modelled in a consistent and homogeneous manner.

Monitoring activities is one of the concepts where meta-models are used. For example, a meta-model for properties associated with software during execution is presented in [16] to ensure the quality of software and its dynamic adaptation after deployment. Indeed, these properties provide a means to assess and improve the resilience of software through the adaptation and anticipation of abnormal situations. The instances of this meta-model are in this case a model with monitoring properties adapted to the target software. In [17], the authors propose a meta-model for the monitoring of cyber-physical systems (CPS), particularly sensor and actuator networks which require valid data and good coordination between sensors and actuators for their operation.

B. Sewerage networks business modelling: related works

To help decision makers in collecting the data necessary for interventions and to diversify their data sources, some solutions have been published. For example: in [10], the authors apply deep neural networks to detect the position of manholes, visible on the ground, from a high-resolution image. In [4], to create the cartography of underground networks, researchers use Bayesian fusion techniques to combine hypotheses extracted from the Ground Penetration Radar (GPR), the spatial location of surveyed manholes, as well as the expectations from the statutory records. However, none of these approaches have been submitted along with a data model.

The work in [18] is an attempt to design a business data model for sewerage networks to build the digital map of the sanitation network for a municipality in Algeria and to contribute to its efficient management. The authors propose, from the inventory of the various available data sources, a conceptual data model on which the necessary objects for the management and their relationships are listed.

At the initiative of the Aquitaine region and a public interest group (Planning and risk management), the Commission of Data Validation for Spatialized Information (COVADIS) has published a data standard for drinking water and sewerage networks intended for French municipalities [19]. The committee presented a class diagram describing the minimum and necessary data to be used by the actors participating in the management of these networks (municipalities, PEIC[1], delegates public services, etc.) for the purpose of a simple data exchange between them.

Since this standard describes the minimum necessary, but sufficient, data for the management of the water networks, and since it is also a standard to be adopted at a nationwide level, we adopt it in our work as a business data model. We present in the following the subpart of the COVADIS class diagram related to the sewerage networks (Figure 3).

It is composed of 4 main classes: Node, Pipe, Reparation and Meta-data whose attributes and related possible values are listed:

- Nodes: represented geometrically by points, they illustrate apparatus (valve, counter etc.) or manholes.
- Pipes: represented geometrically by lines, they are classified into several categories: wastewater, rainwater etc. Each of the pipes has two end Nodes.
- Repairs: geometrically represented by points, they refer to interventions made in Nodes or Pipes.
- Metadata: are data used to qualify the information of the classes Nodes and Pipes. Namely the name of the source, the date of the last update, the reliability of the year of installation and the quality of the geolocation within respect to the 2012 decree [7], which defines 3 precision classes: less than 40 cm, in the range 40 cm and 1.5 m and greater than 1.5 m.

[1]Public Establishment for Inter-municipal Cooperation.

C. Sewerage networks and Big Data

Nowadays, we are witnessing an intensive production of data. Every day, a huge amount of data is produced by companies, on social networks, during transactions or through sensors, that conventional computer tools can no longer process and analyse. The research work around these masses of data, also called Big Data, is a response to these obstacles.

On the other hand, and despite the tiny amount of the available data on sewerage networks compared to the Big Data, they share two important characteristics:

- The multitude of data sources.
- The heterogeneity of the data.

The research works in the domain of the underground networks is mainly confidential [8]. Therefore, the number of publications related to Big Data is more important compared to sewerage networks. To our knowledge, there is no data model for data sources of underground networks. To fill this gap, and since these two characters of the multitude and heterogeneity of the sources have already been examined in Big Data (see for example [20] or [21]), we have chosen to draw inspiration from the solutions proposed in this field.

The available Big Data systems and platforms are not identical, as they come from multiple providers whose vision is not uniform. In [22] using model-driven engineering, the authors propose a platform-independent meta-model to describe the structures of data sources involved in feeding these large volumes of data, thus allowing programmers to create applications compatible with various products. Given that Big Data include three heterogeneous formats: Unstructured, SemiStructured and Structured [23], the authors propose a Meta-Modeling of the three types of data sources as follows:

- Structured: data whose set of possible values are determined and known in advance, such as relational databases.
- Semi-structured: data that have not been organized into a specialized repository. However, they contain meta-data information, which help their exploitation, for example e-mails.
- Unstructured: data represented or stored without a predefined format.

V. CONTRIBUTION

A. Meta-model for sewerage networks data sources

Our target is to propose a generic model for data sources, with the aim of using fusion approaches to combine their data. Therefore, on the one hand, our solution should encompass the available approaches for collecting data. On the other hand, the data sources are diverse and may change over time. An exhaustive modelling of data sources and their possible relationships is not a generic solution. Figure 4 illustrates our meta-model.

For a better understanding, four viewpoints of this meta-model are presented in the following paragraphs.

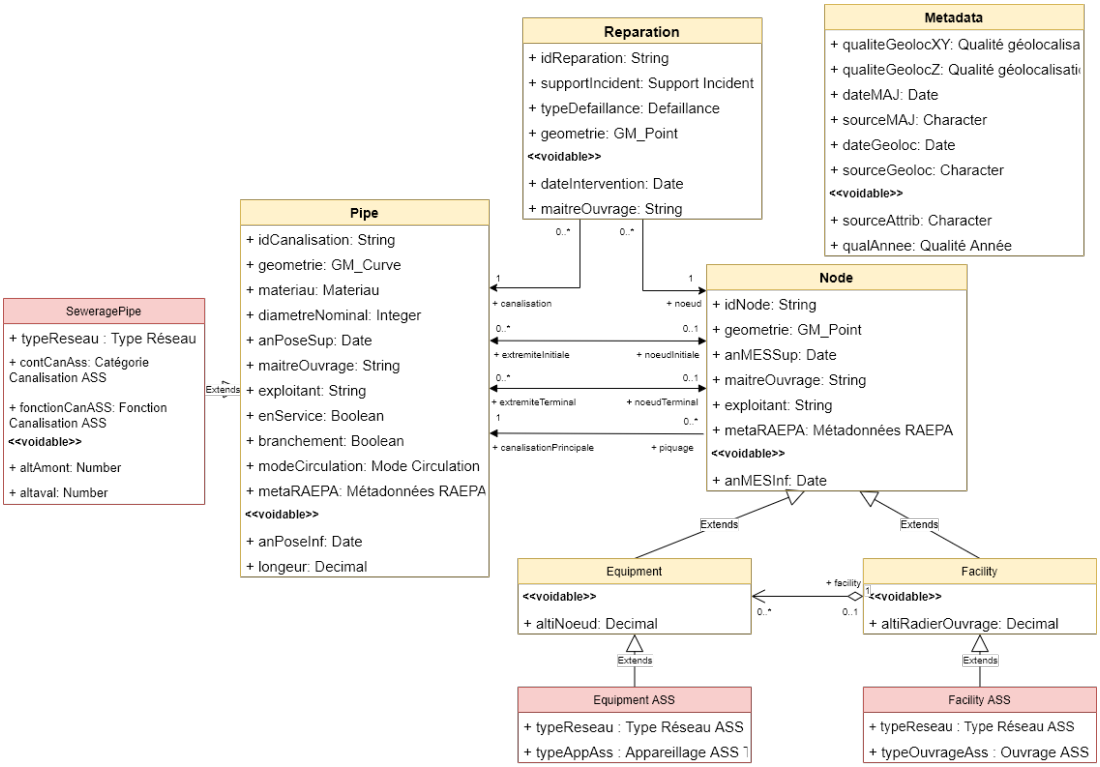

Fig. 3: COVADIS sewerage networks business model [19]

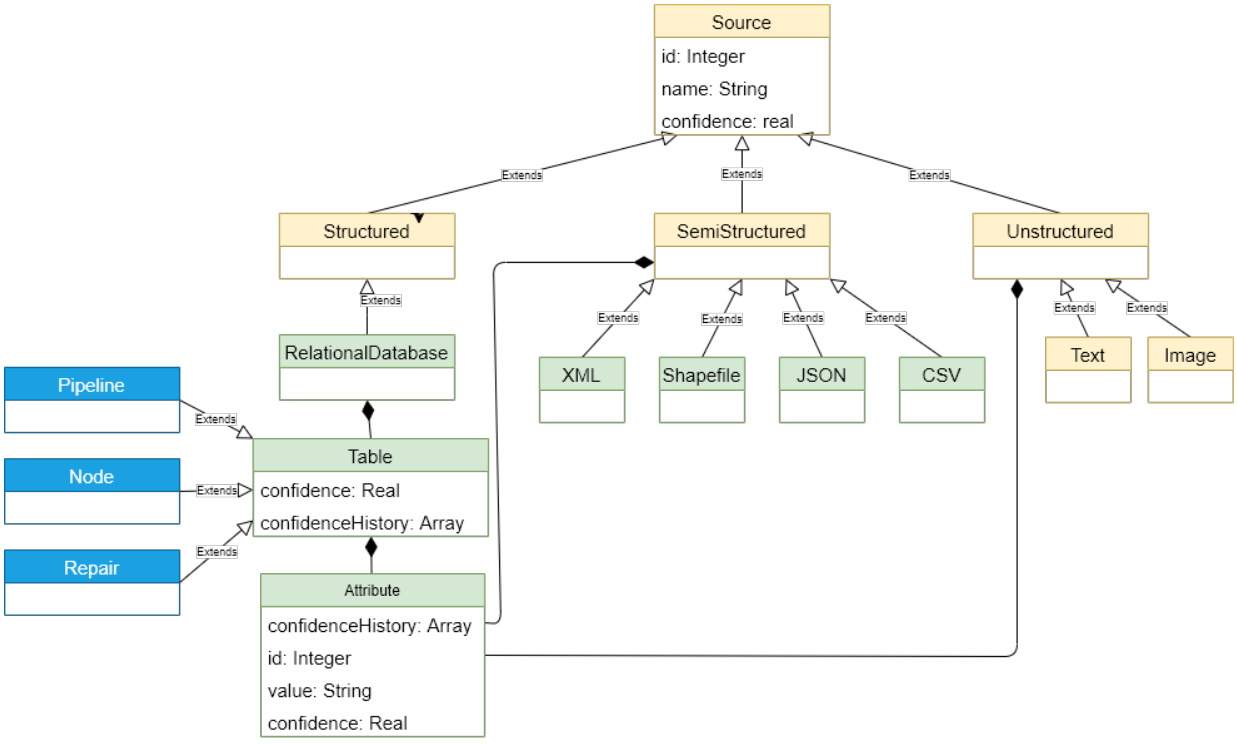

Fig. 4: Meta-model for sewerage networks data sources

B. Data sources viewpoint

We summarise in Figure 5 the data sources viewpoint where the main entity "source" characterises any entity capable of providing data, information or knowledge about sewerage networks. Interpretation of the structuring aspect of data sources is as follows:

- Unstructured sources: whose formats require significant pre-processing before extracting business data about a network. This step generally produce semi-structured data about the networks. For example, CSV file for the location of sewer manholes detected from images.
- SemiStructured sources: whose formats require simple pre-processing before extracting business data about a network. For example, parsing CSV or XML files.
- Structured sources: represent relational databases that directly provide business data about a network. Generally, these data are provided by the official managers of sewerage networks.

C. Attributes viewpoint

The attributes and their relationships with data sources are highlighted in this viewpoint (Figure 6). Each attribute is an entity identified by a name and possesses a String value representing a semantic data. The aggregation of attributes by data sources is encapsulated within the "TWithAttributes" entity. Since it represents, in the case of structured sources, the different attributes within the table of a relational database. Thus, the entity "Table" is connected to this entity. As for semi-structured and unstructured sources, it includes the pre-processing operations, defined by instances of this meta-model, to extract attribute values about sewerage networks. Moreover, data source path is handled within "TWithPath" entity.

D. Confidences viewpoint

Data imperfection is considered in this viewpoint (Figure 7) by allowing confidence attributes to each source, table (representing an object of the sewerage network) and to each attribute characterizing this table (the object). This means that:

- Each source has a confidence value that indicates the reliability or the certainty of the information it provides. This metric can be modelled by the various available mathematical tools, such as probabilities.
- For the components or objects of sewerage networks, this value represents the uncertainty regarding their existence. Indeed, it is possible, for example, for a pipe or a manhole to be represented on a map by mistake.
- As for attributes, the confidence is related to the confidence of the data sources providing them.

Moreover, objects and attributes default confidence values are those associated with their sources. However, this does not imply not having different confidence values later on. For example, an approach identifying manhole covers from images would define the existence confidence of the detected object as the precision of its detection. Meanwhile, the attribute position of the detected object may have a different value, since the detection and the computation of its position are two separate operations. To keep track of the previous data fusion operations, the confidence history is stored too. This would allow knowledge propagation when new information is collected. Several theories support this propagation such Probability theory and Belief theory [24].

E. Business model viewpoint

In our meta-modelling, we distinguish Business data, characterizing the sewerage networks' components, from data about the sources, from which business data are extracted, such as images, documents, calls for tenders, etc.

Figure 8, illustrates this business viewpoint where we adopted the COVADIS (IV-B) standard classes that inherit the properties of the Table entity. For genericity purposes, other business models may easily replace it, provided that the appropriate connections are respected.

VI. Use case

A. Data aggregation for data fusion purposes

Through our proposed meta-model, we aim to perform the fusion of data coming from different sources. In fact, successful fusion operations require a global and complete knowledge about the available data and their sources. To achieve this goal, heterogeneous data, collected from various sources, should be aggregated into a single entity.

According to the data sources viewpoint of our meta-model (Figure 5), we classified the sources as structured, semi-structured and unstructured. Aggregating the attributes' values according to the specificities of each single data-source, is considered in the attributes viewpoint of the meta-model within the TWithAttributes and TWithPath traits (figure 6). Figure 9 depicts the main steps to reach data aggregation from these different sources for fusion purposes.

The first step consists in extracting data from unstructured data sources through automatic and semi-automatic approaches. This usually requires the use of additional technics related to computer vision, image processing and data mining. Detection of the aerial elements of buried networks from an image is one concrete example. The output of this step is recorded in a semi-structured format, such as flat-file format CSV or JSON.

The second step focuses on the necessary transformations that should be applied to the semi-structured data, together with the available structured data, in order to adapt them to the targeted business data model such as the COVADIS model. These operations of parsing, importing and transforming should be defined for each semi-structured data source and are implemented through what we call a *transformer* for the semi-structured data sources and an *adapter* for the structured data sources. These operations should also include a model manager and an intermediate data visualizer.

Once the data of each single source are pre-processed, the data of all sources could be aggregated allowing to visualize

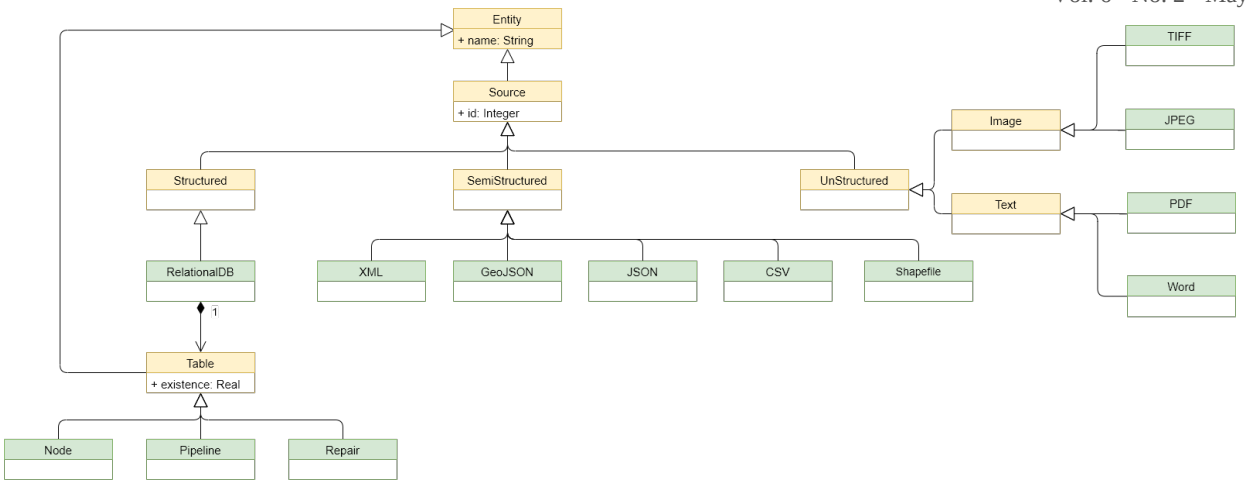

Fig. 5: Data sources viewpoint

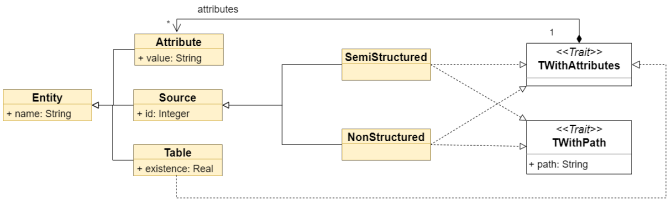

Fig. 6: Attributes viewpoint

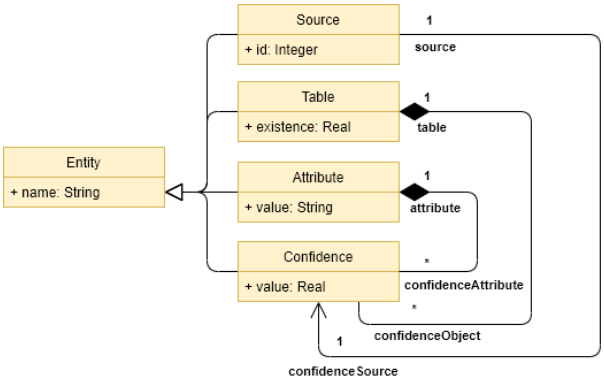

Fig. 7: Confidences viewpoint

and inquiry them in a unified manner. At this step, the dataset is ready and the fusion operations can be conducted.

To demonstrate this workflow, we conducted an experiment of aggregating data on the sewerage network of Prades-le-Lez, a town located on the outskirts of the city of Montpellier in France, from 3 different sources. Our first source is the open data database of Montpellier Metropole Médieterrannée, which is a semi-structured source that we note the 3M source. The two other sources are a High-Resolution image database and Google Street View images, which are both unstructured sources that we note respectively HR and GSV sources. To conduct our experiment, we implemented our demonstration using the Moose platform, a software analysis platform.

In the following paragraphs, we will briefly describe the Moose platform and its features. We will then present our different sources and their specific data extraction methods, as well as the data aggregation we performed in the case of a single data source and multiple data sources. Finally, we will report and discuss the obtained results.

B. Moose

Moose [6], [25] is an open source platform for software and data analysis, currently based on Pharo [26] a pure object-oriented programming language. It is intended for researchers, engineers, software architects as well as tool builders. The platform is mainly used for software analysis through multiple available mechanisms and features:

- Importing and meta-meta-modelling, since the first step in the process of analysis is the generation of a model of a given target system,
- Parsing, which provides a fluent interface for easy construction,
- Analysis, which provides a rich interface for querying models,
- Visualization through graphs and charts,
- Browsing, which enables the analyst to browse any model.

Moose is also designed to help the programmer build his own tools. This is achieved by the means of several engines through which he can control and customize the complete analysis workflow. In particular:

- build new importers for new data sets,
- define new models to store the data, and
- create new analysis algorithms and tools such as: complex graph visualizations, charts, new queries, or even complete browsers and reporting tools altogether.

Moose was started in the context of the FAMOOS European project (1996-1999), a project focusing on methods

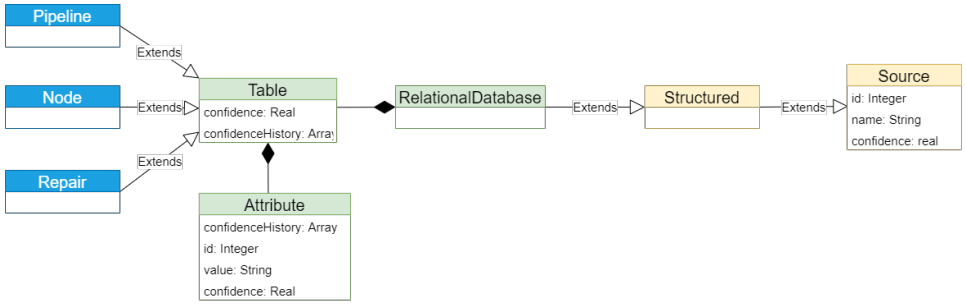

Fig. 8: Business model viewpoint

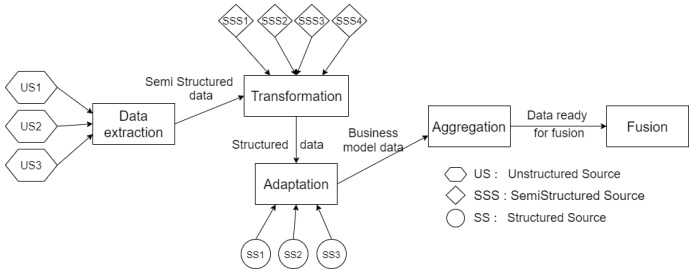

Fig. 9: The main steps towards data fusion

and tools to analyze and detect design problems in object-oriented legacy systems, and to migrate these systems towards more flexible architectures. Since then, Moose continued to enhance and integrate new features and tools. Its development is currently supported by multiple research groups, startups and contributors.

C. Aggregation of a single semi-structured data-source

Our first data source is the Open data database of Montpellier Méditerranée Métropole (3M). Through its website [27], this intercommunal structure provides to the public numerous open data files concerning various subjects: transport, finance, environment, etc. The wastewater network datasets are part of these files and include information about nodes and pipes such as the positions of nodes and the dimensions of pipes. We used the dataset related to the position of nodes of the town of Prades-le-Lez town, that we exported as a CSV file.

To get data from a CSV source, we defined an importer MsgMonitoringCSVImporter which takes the path for the CSV file as parameter. The CSV format is popular, thus in Moose platform a default parser for CSV data sources is already defined, the PPCommaSeparatedParser. However, since this parser cannot handle, among other things, white spaces in the values, we have extended it and defined our proper parser: MsgMonitoringCSVParser which takes as input the CSV file content as String value.

Through a stream operation, the importer reads data from the source, handles it to the parser and then passes the resulted parsed data to the CSV model manager: MsgMonitoringCSVModelManager. Since there is no meta-model in Moose for CSV files, the manager designates the parsed output

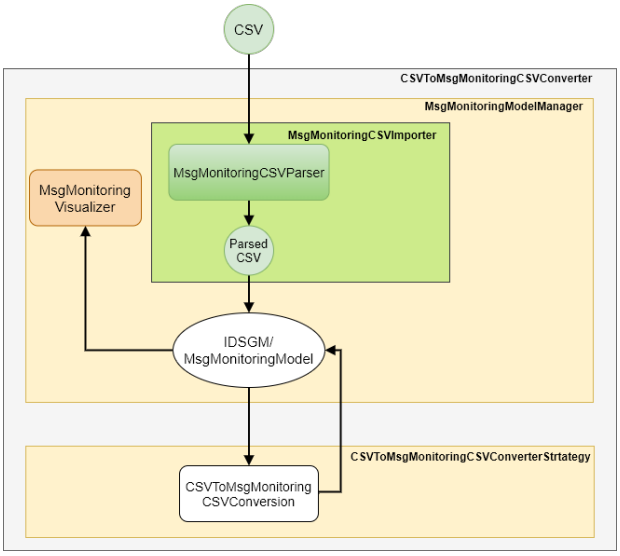

Fig. 10: Workflow for the aggregation of a single data source (CSV)

as Intermediate Data Source General Model (IDSGM). Figure 10, illustrates this process.

To visualize the different elements of the MsgMonitoringCSV, we defined the visualizer MsgMonitoringCSVVisualizer whose results are shown in the figure 11. The left side of the screen shows the used source (file ”source1.csv” containing data of Prades-le-Lez) and the right side displays a browser to navigate through the different attributes and their values from the source.

D. Our unstructured data-sources

High resolution images: Extracting data from this source is the result of the work described in [10]. The authors used deep convolutional neural network to detect and extract manhole covers from a very high resolution aerial images, which were purchased specifically for their study from a specialized company. This work was conducted on Prades-le-Lez and Gigean, two towns of the 3M (Montpellier Méditerranée Métropole). The Prades-le-Lez dataset was composed of 6 20000 $\times$ 20000 pixel georeferenced images, and the Gigean dataset was composed of one image of 17749 $\times$ 18361 pixels.

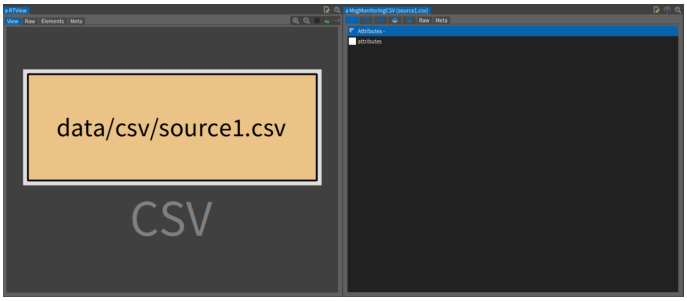

Fig. 11: visualization of the aggregation of a single data source (CSV)

Data augmentation techniques such as rotation, translation and horizontal flip, were used on a small portion of the available datasets in order to train the CNN. The predictions were performed on the remaining unlabeled portion. The validation phase was done manually by an expert. The authors were able to detect 49% of ground truth manholes with a precision of 75%.

For the purpose of our experiment, we used the CSV semi-structured format to store the output of the manhole detection from high resolution images, and thus we used the workflow described in Figure 10 to aggregate the data from this source in Moose.

Google Street View Images: Considering that in reality, a large part of the manhole covers are located on roads and streets, we propose as a second unstructured source for our experiment, to detect manhole covers from Google Street View Images. Indeed, they provide an interactive panorama of the majority of streets in France, and many countries around the word. To get the positions of manhole covers of Prades-le-Lez, we proceeded by collecting all the images of Prades-le-Lez corresponding to the nodes positions provided by our first, official data source (3M). We proceeded this way in order to reduce the number of unnecessary images to be treated.

The following 3 main steps summarize our data collection process:

- Using Google Static Street View API, we were able to collect images corresponding to the positions provided by 3M.
- To detect manhole covers, we used the You Only Look Once (Yolo) object detection algorithm [28], that we trained on thousands of manhole covers.
- We exported the positions of the detected manhole covers along with the precision of the detection to an XML file. Figure 12, illustrates an example of a detected manhole cover from a Google Street View image.

We used XML as a semi-structured storage format for this data source to demonstrate the usefulness of the proposed aggregation of data from multiple semi-structured formats. Therefore, we defined an XML workflow (Figure 13) similar to the CSV workflow (Figure 10). In other words, we extended the default Moose Parser for XML PPXmlParser to MsgMonitoringXMLParser, and defined the different components

Fig. 12: Example of manhole covers' detection using Google Street View images

necessary for the workflow: the importer MsgMonitoringXMLImporter, the model manager MsgMonitoringXMLModelManager, the conversion XMLToMsgMonitoringXMLConversion, the strategy XMLToMsgMonitoringXMLConversionStrategy and the visualizer MsgMonitoringXMLVisualizer. The visualization for the different sources is managed by the MsgMonitoringVisualizer.

E. Aggregation of multiple data sources.

To aggregate data from multiple sources in Moose, a coordinator is necessary. Therefore, we defined ModularIDGSMToMsgMonitoringConverter, which contains the different convertors prescribed for each data source as modules. The role of this coordinator is to choose automatically for each data source, the corresponding converter. The coordinator receives data sequentially from different data sources, then for each iteration, the data-source associated module is identified and used to trigger the workflow explained bellow. That is to say, generating the IDSGM model from the parsed data using the importer, and then aggregating data to the associated model from the MsgMonitoringModel (Figure 13).

Figure 14, shows the visualization of our 3 semi-structured data-sources in Moose, two CSVs (the 3M and HR sources) and one XML (the GSV source), managed by the MsgMonitoringVisualizer.

To finalize our demonstration, we developed an API for the analysis, interaction and monitoring of the aggregated data provided by the different sources, through a series of queries and requests. To achieve this goal, we based our API on the design patterns Visitor and Strategy. Our meta-model has a hierarchical structure starting from the global entity source

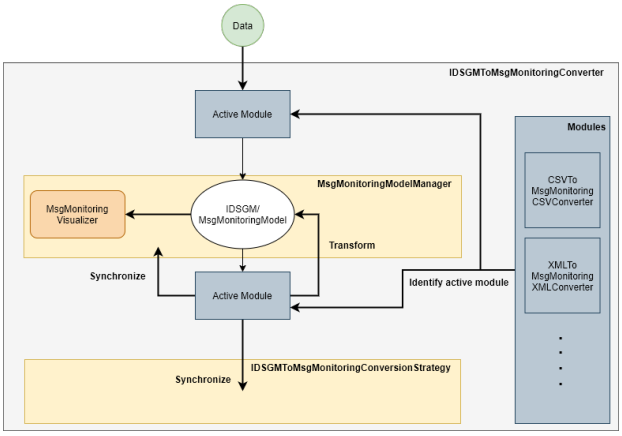

Fig. 13: Workflow for the aggregation of multiple data-sources (CSV and XML)

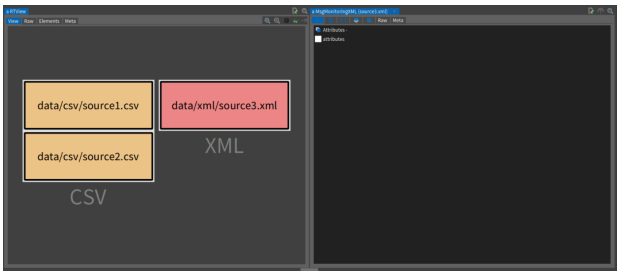

Fig. 14: Visualization of the aggregation of three data-sources

to the single attribute values. Consequently, we defined three levels of the analysis:

- Source level analysis: where the available sources and their specific information like their providers and the associated confidence values are listed. We used the Visitor design pattern to implement this level.
- Object level analysis: using also the Visitor design pattern, this level lists the different available objects for each source. For example, the 3M source provides data about nodes and pipes, whereas the GSV source produces data about nodes only.
- Attribute level analysis: this level is more sophisticated, since it processes data sources in their deep details, the attributes and their values. Thus, besides the Visitor design pattern we used Strategy which allows defining for each data-source, the specific required operations for their analysis and monitoring. For example, for a CSV data source, and in addition to the elementary requests of attributes values, the user may request the type of separator, the available attributes, the distribution of attributes values, the line at a specific index, the first n lines, lines between two indexes, etc.

F. Results and discussion

Table Ia, illustrates a concrete example of the source level analysis where we list the three sources used in this experiment and their confidence values. This may allow the final user to make conclusions for decision-making or conduct some reasoning to understand the influence of these values on the confidence of other inter-related data, or infer them in deeper levels of analysis. As we mentioned before, the perspective of our work is to automate this reasoning task for fusion purposes. In the current experiment, we can explain the confidence values we assigned to the sources as follows :

- The 3M source is a priori reliable, as it was provided by the official managers of the network. However, on the one hand, data are incomplete as not all of the attributes and their values are listed in the attribute tables. On the other hand, considering that sewerage networks evolve in time, there is no guarantee that this source has continuously updated its data during its years of service. Therefore, taking into consideration all of these facts, a subjective confidence value of 0.9 was assigned to this source by a domain expert.
- We assigned to the HR source the value of 0.75 as an objective confidence. It corresponds to the ”precision” metric of the algorithm of the detection of manholes covers, namely 75%.
- The Google Street View Images: when we collected Google Street View images we used only the node positions provided by 3M. However, not all those positions were associated to an image. In fact, among the 799 positions that we used, we found only 763 corresponding images, more than half of them didn’t contain a manhole cover, which is due to the presence of diverse objects on the street such as cars hiding manholes, machine learning model failing to identify manholes covers, also considering that not all the node positions provided by 3M represent manholes covers but they can represent junctions of underground pipes. Therefore, based on this information we assigned objectively the value 0.6 (451 / 763) as the confidence of this source.

At the object level analysis, we could display the 1862 collected nodes from the 3 sources: 799 nodes from the 3M source, 451 nodes from the GSV source and 612 from the HD images source. Even if we used the node positions provided by 3M to collect data from the GSV source, only 451 of the 799 nodes have been detected.
We aggregated data from the 3M, the GSV and finally the HR source respectively. We categorized the nodes at the attribute level analysis, as appearing in the 3 sources, in 2 from the 3 sources or only in one source. This information is important in the process of data combination to confirm the positions of the nodes previously available in our records and to detect potential nodes that were not identified.
Table Ib. shows examples of nodes where the attributes and their confidence values are listed. It is to be noted that the confidence values we assigned to the attributes represent the precision value of the manhole cover detection from the images; which explains the differences between the confidence values of the nodes in the same source for

TABLE I: Examples of queries using the three sources

(a) Confidence values of the sources

Sources	CSV_3M_source	XML_GSV_source	CSV_HR_source
Confidence	0.90	0.60	0.75

(b) Confidence values of the attribute position

CSV_3M_source			XML_GSV_source			CSV_HR_source		
ID	Position	Confidence	ID	Position	Confidence	ID	Position	Confidence
613	3.8632613,43.6916825	0.90	78	3.8632613,43.6916825	0.75	12	3.8632613,43.6916825	0.76
622	3.8642639,43.7010732	0.90	80	3.8642639,43.7010732	0.58	52	3.8642639,43.7010732	0.80
623	3.8660806,43.6890539	0.90	45	3.8660806,43.6890539	0.98	32	3.8660806,43.6890539	0.84
602	3.8653489,43.6939405	0.90	14	3.8653489,43.6939405	0.96			
617	3.8629708,43.6884132	0.90	19	3.8629708,43.6884132	0.94			
						66	3.8734563,43.6932746	0.82
						207	3.8646406,43.7005578	0.80

the HR and the GSV sources. Since the 3M sources is a semi-structured one, there had been no need to the extraction step and then all the nodes inherited the confidence value of their source.

The first 3 lines of the table represent 3 nodes appearing in the 3 sources (corresponding to 9 detected nodes from the 3 sources). The 2 following lines are nodes appearing in both the 3M and the GSV sources, which allows to confirm their positions. The last 2 lines represent nodes detected only by the HR source. In this case, they may represent an extension of the network that is still not reported by the official source, thus really missing data, or they may be simply false positive cases extracted using the HR source.

Although, the results of the detection from the high resolution images and Google Street View images are far from perfect, this example demonstrates that they can be used to confirm and update the confidence value of the attributes of the 3M source. Such images can also be used to create a fourth dataset where the objects appearing in only one source (Section VI-F) would have low confidence values and those in common would have a fused confidence value.

VII. Conclusion

In this work, we proposed a meta-model for data sources related to sewerage networks inspired from the field of Big Data. The main objective is to perform a multi-source data fusion on sewerage networks to make a more complete dataset available to the decision makers, taking into account data imperfections. Our proposition considered the three important aspects of i) structure of the data sources: structured, semi-structured and unstructured, ii) associated confidences at multiple levels and iii) genericity related to the business domain. As a first step towards sewerage networks data fusion and through our meta-model, we implemented this very initial step in Moose, a platform for software and data analysis. As a concrete example, and to show that our meta-model is generic and able to encompass the various sources and approaches for the collection of data about sewerage networks, we used one semi-structured source and two unstructured sources with their appropriate data extraction processes. The results enabled the listing, the visualization and the analysis of the aggregated data which suffer in most cases from imperfections. We are currently conducting some experiments on spatial sewerage data fusion that we expect to enhance by proposing a data fusion approach with uncertainty management allowing to cop with the multi-source data imperfections. A validation example was provided in the previous section. It should however be noted that additional case studies are necessary to insure the genericity of our proposition.

VIII. Acknowledgments

This work was carried out within the framework of the CIFRE-France / Morocco Program.

References

[1] L. Bernold, L. Venkatesan, and S. Suvarna, “A multi-sensory approach to 3-d mapping of underground utilities,” *NIST SPECIAL PUBLICATION SP*, pp. 525–530, 2003.

[2] ASTEE, “Gestion patrimoniale des réseaux d’assainissement,” https://www.astee.org/, 2015, accessed 01 August 2020.

[3] (2015) Open platform for french public data. https://www.data.gouv.fr/. Accessed 01 august 2020.

[4] H. Chen and A. G. Cohn, “Buried utility pipeline mapping based on multiple spatial data sources: A bayesian data fusion approach,” in *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[5] Y. Bel-Ghaddar, A. Seriai, A. Begdouri, C. Delenne, N. Chahinian, and M. Derras, “Combining model-driven engineering and sewerage networks: towards a generic representation,” in *CiST’20 - 6th IEEE Congress on Information Science and Technology*, Agadir - Essaouira (postponed to June 2021), Morocco, Dec. 2020.

[6] S. Ducasse, T. Gîrba, M. Lanza, and S. Demeyer, “Moose: A collaborative and extensible reengineering environment.” in *Tools for Software Maintenance and Reengineering*. Citeseer, 2005.

[7] (2012) Legifrance. https://www.legifrance.gouv.fr/. Accessed 01 august 2020.

[8] M. Hafsi, P. Bolon, and R. Dapoigny, “Detection and localization of underground networks by fusion of electromagnetic signal and GPR images,” in *Thirteenth International Conference on Quality Control by Artificial Vision 2017*, H. Nagahara, K. Umeda, and A. Yamashita, Eds., vol. 10338, International Society for Optics and Photonics. SPIE, 2017, pp. 7 – 14.

[9] D. Boller, M. Moy de Vitry, J. D. Wegner, and J. P. Leitão, “Automated localization of urban drainage infrastructure from public-access street-level images,” *Urban Water Journal*, vol. 16, no. 7, pp. 480–493, 2019.

[10] B. Commandre, D. En-Nejjary, L. Pibre, M. Chaumont, C. Delenne, and N. Chahinian, "Manhole Cover Localization in Aerial Images with a Deep Learning Approach," *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42W1, pp. 333–338, May 2017.
[11] G. Kabir, S. Tesfamariam, J. Hemsing, and R. Sadiq, "Handling incomplete and missing data in water network database using imputation methods," *Sustainable and Resilient Infrastructure*, vol. 5, no. 6, pp. 365–377, 2020.
[12] Y. Belghaddar, N. Chahinian, A. Seriai, A. Begdouri, R. Abdou, and C. Delenne, "Graph convolutional networks: Application to database completion of wastewater networks," *Water*, vol. 13, no. 12, p. 1681, 2021.
[13] Object Management Group. (2006) Meta object facility (MOF) 2.0 core specification. https://www.omg.org/spec/MOF/2.0/. Accessed 1 august 2020.
[14] A. R. Da Silva, "Model-driven engineering: A survey supported by the unified conceptual model," *Computer Languages, Systems & Structures*, vol. 43, pp. 139–155, 2015.
[15] J. M. Gascueña, E. Navarro, and A. Fernández-Caballero, "Model-driven engineering techniques for the development of multi-agent systems," *Engineering Applications of Artificial Intelligence*, vol. 25, no. 1, pp. 159–173, 2012.
[16] A. Bertolino, A. Calabrò, F. Lonetti, A. Di Marco, and A. Sabetta, "Towards a model-driven infrastructure for runtime monitoring," in *Software Engineering for Resilient Systems*, E. A. Troubitsyna, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 130–144.
[17] T. B. la Fosse, Z. Cheng, J. Rocheteau, and J. M. Mottu, "Model-driven engineering of monitoring application for sensors and actuators networks," in *2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 2020, pp. 553–560.
[18] C. Abdelbaki and M. Zerouali, "Modélisation d'un réseau d'assainissement et contribution a sa gestion a l'aide d'un système d'information géographique-Cas du chef lieu de commune de Chetouane-wilaya de Tlemcen Algérie," *LARHYSS Journal P-ISSN 1112-3680/E-ISSN 2521-9782*, no. 10, 2012.
[19] COVADIS. (2019) Standard de données réseaux d'AEP & d'assainissement, version 1.2. http://www.geoinformations.developpement-durable.gouv.fr/. Accessed 01 August 2020.
[20] X. L. Dong and D. Srivastava, "Big data integration," in *2013 IEEE 29th international conference on data engineering (ICDE)*. IEEE, 2013, pp. 1245–1248.
[21] A.-C. Boury-Brisset, "Managing semantic big data for intelligence." in *Semantic Technologies for Intelligence, Defense, and Security*, 2013, pp. 41–47.
[22] A. Erraissi and A. Belangour, "Data sources and ingestion big data layers: meta-modeling of key concepts and features," *International Journal of Engineering & Technology*, vol. 7, no. 4, pp. 3607–3612, 2018.
[23] A. Oussous, F.-Z. Benjelloun, A. A. Lahcen, and S. Belfkih, "Big data technologies: A survey," *Journal of King Saud University-Computer and Information Sciences*, vol. 30, no. 4, pp. 431–448, 2018.
[24] A. Appriou, *Uncertainty theories and multisensor data fusion*. John Wiley & Sons, 2014.
[25] Moose. https://moosetechnology.org. Accessed 1 august 2020.
[26] Pharo. https://pharo.org/. Accessed 1 august 2020.
[27] Montpellier Méditeranée Métropole. Open data. https://data.montpellier3m.fr/. Accessed 1 august 2020.
[28] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *ArXiv*, vol. abs/1804.02767, 2018.