

A Data Mining Approach for a Dynamic Development of an Ontology-Based Statistical Information System

Mohamed Hachem Kermani* 

National Polytechnic School of Constantine-Malek Bennabi,
Constantine, Algeria
Laboratoire d'Informatique Repartie (LIRE) Laboratory, Constantine,
Algeria
E-mail: hachem.kermani@enp-constantine.dz

Zizette Boufaida 

University of Constantine 2 - Abdelhamid Mehri, Constantine, Algeria
Laboratoire d'Informatique Repartie (LIRE) Laboratory, Constantine,
Algeria
E-mail: zizette.boufaida@univ-constantine2.dz

Amel Lina Bensabbane 

University of Constantine 2 - Abdelhamid Mehri, Constantine, Algeria
Laboratoire d'Informatique Repartie (LIRE) Laboratory, Constantine,
Algeria
E-mail: amel.bensabbane@univ-constantine2.dz

Besma Bourezg 

University of Constantine 2 - Abdelhamid Mehri, Constantine, Algeria
Laboratoire d'Informatique Repartie (LIRE) Laboratory, Constantine,
Algeria
E-mail: besma.bourezg@univ-constantine2.dz


ABSTRACT

This paper presents a dynamic development of an ontology-based statistical information system supporting the collection, storage, processing, analysis, and the presentation of statistical knowledge at the national scale. To accomplish this, we propose a data mining technique to dynamically collect data relating to citizens from publicly available data sources; the collected data will then be structured, classified, categorized, and integrated into an ontology. Moreover, an intelligent platform is proposed in order to generate quantitative and qualitative statistical information based on the knowledge stored in the ontology. The main aims of our proposed system are to digitize administrative tasks and to provide reliable statistical information to governmental, economic, and social actors. The authorities will use the ontology-based statistical information system for strategic decision-making as it easily collects, produces, analyzes, and provides both quantitative and qualitative knowledge that will help to improve the administration and management of national political, social, and economic life.

Keywords: statistical information, data mining, dynamic development, ontology-based information system, statistical information system, strategic decision-making

Received: May 1, 2022
Accepted: March 9, 2023

Revised: March 3, 2023
Published: June 30, 2023

***Corresponding Author:** Mohamed Hachem Kermani
 <https://orcid.org/0000-0002-2315-6951>
E-mail: hachem.kermani@enp-constantine.dz



All JISTaP content is Open Access, meaning it is accessible online to everyone, without fee and authors' permission. All JISTaP content is published and distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>). Under this license, authors reserve the copyright for their content; however, they permit anyone to unrestrictedly use, distribute, and reproduce the content in any medium as far as the original authors and source are cited. For any reuse, redistribution, or reproduction of a work, users must clarify the license terms under which the work was produced.

1. INTRODUCTION

The development of ontology-based information systems is argued when applied to the analysis, conceptual modeling, design, and re-engineering of complex information systems. Ontologies serve as a theoretical foundation for the conceptual model, which is used to design and develop the entire information system. Ontologies enable specialists, software agents, systems, and services to share a common understanding of a domain. Ontologies have also influenced research and development in other areas of computational intelligence, resulting in a plethora of hybrid information systems. They have introduced a new way of thinking about, researching, and developing information systems, which, in conjunction with other technologies, has resulted in the emergence and development of intelligent information systems, such as statistical information systems.

Indeed, a system that provides statistical information aids in the preparation, execution, and assessment of decisions and actions pertaining to a complete or partial system, such as a community and its inhabitants, establishments and their operations, or a company and its staff, customers, and vendors with their activities. The statistical information system is intended to acquire, retain, handle, scrutinize, display, and obtain information. These responsibilities are comparable to those of data mining, which refers to the examination of a vast pre-existing database to create new insights. A variety of data mining techniques have been formulated and utilized in recent times, including association, classification, clustering, decision tree, prediction, and Neural Networks (Kermani & Boufaïda, 2022).

Obviously, these methodologies are interconnected with the operation of ontologies, and it can be challenging to differentiate the boundary between data mining and ontology. According to Plirdpring and Ruangrajitpakorn (2022), an ontology is a precise and explicit account of the concepts within a specific domain, including the properties and attributes associated with each concept. The collection of individual instances of these concepts with an ontology constitutes a knowledge base. This research presents a data mining strategy for the real-time construction of an ontology-based statistical information system that arranges, sorts, and categorizes information concerning citizens, enabling government, economic, and social authorities to access accurate statistical data.

To accomplish this, we must 1) dynamically collect (i.e. automatically and continuously) all data pertaining to citizens from publicly available data sources (i.e. civil status

files, government databases, etc.); 2) structure, classify, and categorize knowledge about citizens before dynamically integrating it into the ontology; and 3) develop an intelligent ontology exploitation platform that provides authorities with quantitative and qualitative information about the country's social and economic situation. The rest of this paper is organized according to the following. Section 2 provides an overview of research that is related to our approach. Section 3 presents our proposal, which is a data mining approach for the dynamic development of an ontology-based statistical information system. Section 4 presents a software application and experimentation of the proposed approach. Section 5 presents a discussion. Finally, Section 6 concludes the paper and suggests directions for future research.

2. RELATED WORK

In our proposed approach, we investigate data mining techniques, ontologies, and statistical information systems development. These aspects have been the subject of numerous works in the literature, which we will discuss below.

2.1. Statistics Data Mining Approaches

The discipline of data mining encompasses computer science and statistics, and seeks to intelligently extract information from data sets and transform it into a comprehensible format for further use. Various statistical data mining methods have been proposed for this purpose, such as the approach described in a study by Baek et al. (2018). This study used electronic health records to analyze event logs of patients admitted to a South Korean university hospital between January and December 2013, with the aim of identifying factors affecting the length of hospital stays. The authors aimed to develop a system that could assist hospitals in managing inpatient admissions more effectively.

Furthermore, in a study conducted by Fernandes et al. (2019), a data mining approach was used to predict the academic performance of public school students in the capital city of Brazil. The authors initially performed a descriptive statistical analysis to gain insights from the data, which resulted in the creation of two datasets. The first dataset consisted of variables obtained before the start of the school year, while the second contained academic variables collected two months after the semester began. The authors developed classification models based on Gradient Boosting Machine for each dataset, to predict the academic outcomes of student performance at the end of the

school year. The study revealed that 'grades' and 'absences' were the most important attributes for predicting end-of-year academic outcomes, but demographic attributes such as 'neighborhood,' 'school,' and 'age' also showed potential as indicators of academic success or failure.

In the same context, another study conducted by Adekitan and Noma-Osaghae (2019) introduced a data mining technique to predict the academic performance of first-year university students. The research examined the correlation between the academic records of students admitted into Covenant University in Nigeria and their performance during their first year using predictive data mining and regression models. The data mining model was applied on KNIME and Orange platforms, and the accuracy of the prediction was verified using regression analysis. The study recorded maximum accuracies of 50.23% and 51.9%, respectively, with R^2 values of 0.207 and 0.232, indicating that students' cognitive entry requirements do not entirely explain their performance during the first year.

Besides this, a research paper authored by Jatav (2018) presented an algorithm for medical diagnosis using predictive data mining. The study focused on predicting diabetes, kidney, and liver disease using a vast number of input attributes. To evaluate the performance of the algorithm, data mining classification techniques such as support vector machine (SVM) and random forest (RF) were used on respective databases for each disease. The study compared the performance of these techniques using metrics such as precision, recall, accuracy, f-measure, and time. The experimental results showed an accuracy of 99.35%, 99.37%, and 99.14% on diabetes, kidney disease, and liver disease, respectively, and concluded that the proposed algorithm was created using the SVM and RF algorithms.

2.2. Approaches for Developing Ontology-Based Information Systems

According to Sowa (2011), all software systems have an implicit or explicit ontology, which means they work with a knowledge base that can be represented using ontologies. Indeed, Nawi et al. (2021) suggest that designing an information system based on a formal ontology is possible, and many works on ontology-based information systems have been proposed in recent years. For instance, Luković et al. (2019) introduced an ontology-based module for the information system ScolioMedIS that uses ontologies for 3D digital diagnosis of adolescent scoliosis. The researchers followed four steps, namely specification,

conceptualization, formalization, and implementation, to develop the OBR-Scolio ontology and the ontology-based module of the ScolioMedIS. They used the Protégé-OWL API to create, edit, delete, and query the ontology. The module was tested on datasets of 20 female and 15 male patients with AIS, and it automatically generated statistical indicators about the frequency and characteristics of spinal curvatures based on the Lenke classification system and Lenke scoliosis types. The results were compared with the analysis of 315 observed patients performed using traditional radiation techniques.

Furthermore, a study by Comas Rodríguez et al. (2019) presented a data management model called OntoSIGOBE that is applied to the Geographic Information System SIGOBE, which is a part of the Business Management System of the Cuban Electric Union (SIGE). This model utilizes a developed domain ontology and a query answering process based on case-based reasoning technique to achieve semantic interoperability between heterogeneous data sources and the query answering process. The results of the study showed that OntoSIGOBE provides flexible and integrated data access in SIGOBE and is highly satisfactory to end users. Similarly, Ledvinka et al. (2019) presented an ontology-based information system for aviation safety data integration that was developed in cooperation with the Civil Aviation Authority of the Czech Republic. The study described the system's design, core technologies, and achieved functionalities, and compared it to other available solutions in the domain. The results showed that by linking the features to ontologies, the system achieved many desired characteristics of a safety information system, providing a systematic and sustainable way to improve aviation safety.

Besides this, Choi and Choi (2019) put forward a security context reasoning approach based on ontology for power IoT-Cloud security service, which presented a suitable framework for power IoT-Cloud environment security service. Furthermore, an effective security mechanism was implemented and tested using attack context scenarios based on a smart meter, a critical component of power systems. Inference rules were created for each attack stage to identify attack paths that exploited smart meter vulnerabilities, and the results showed a high level of attack detection. While most literature focuses on data mining and ontology-based information systems for statistical purposes, we propose a data mining approach for a dynamic ontology-based statistical information system to provide reliable quantitative and qualitative information on citizens to governmental, economic, and social author-

ities, thereby facilitating a comprehensive understanding of any country's socio-economic situation.

3. THE PROPOSED APPROACH

Statistics is the science of learning from data by collecting and analyzing information in order to effectively present the results. Statistics is an important part of how we make scientific discoveries, make data-driven decisions, and make predictions. Statistics are useful when used to improve management and decision making. Therefore, recent research has focused on employing computational techniques from statistics, data mining, and information theory to automatically collect and analyze massive amounts of data. Our proposed data mining approach for the dynamic development of an ontology-based statistical information system (Fig. 1) is part of these research efforts to provide governmental, economic, and social authorities with reliable statistical data in order to improve management and decision making. The proposed approach is based on the Model View Controller (MVC) architecture, with the *Controller* representing a data mining algorithm that dynamically (i.e. automatically and continuously) collects all data pertaining to citizens from publicly available data sources (i.e. civil status files, government databases, etc.), and then structures, classifies, and categorizes these knowledge into the ontology. The *Model* is a national ontology that integrates all knowledge about citizens, and the *View* is an intelligent ontology exploitation platform that

provides authorities with quantitative and qualitative data about the country's social and economic situation.

Our approach to developing a statistical information system based on an ontology uses a data mining method. This method is based on the MVC pattern, an architectural framework that breaks down an application into three major conceptual components: the Model, the View, and the Controller. While originally designed for desktop graphical user interface applications, this pattern gained popularity with web apps and is now widely supported across common languages (Kermani et al., 2021). Our approach follows this pattern by dividing each component according to its adherence to the MVC pattern. The Model, which serves as the central component of our system, is the national ontology. The View, which represents the output of knowledge, is an intelligent ontology exploitation platform. The Controller is a data mining algorithm that dynamically collects, structures, and categorizes citizen knowledge into an ontology. More details on each of the three components of our system are presented in the following sections.

3.1. The Data Mining Algorithm

In the MVC pattern, the controller plays a vital role as it oversees the decision-making process, system logic, and serves as a mediator between the model and the view. In our suggested MVC framework, the data mining algorithm fulfills the controller's duties. It is the critical element of our system that gathers, arranges, and groups all citizen information into the ontology.

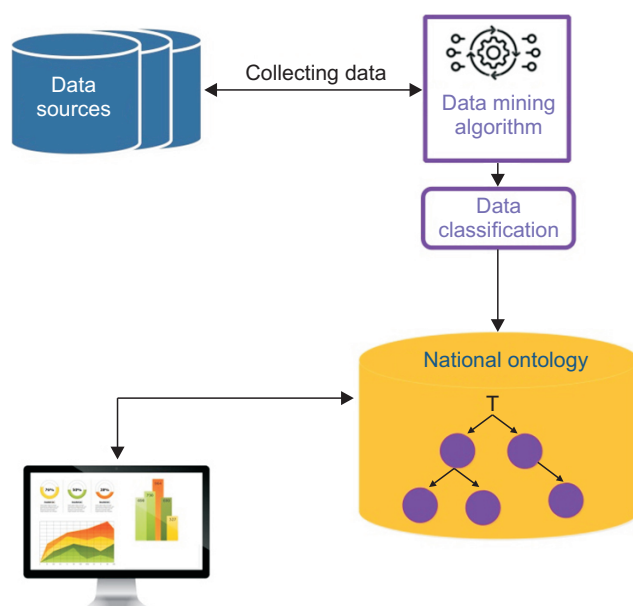


Fig. 1. The ontology-based statistical information system.

The data mining algorithm	
1:	Input: CSF // Civil Statutes Files.
2:	Input: ADB // All Data Bases.
3:	Begin
4:	ID = 1
5:	While (ID ≠ 0) Do
6:	Locate (ID, CSF) // Select ID from Civil Statutes Files.
7:	Recuperate Data () // Retrieve all data related to the selected ID.
8:	Search (ID, CSF) // Searching ID in CSF where ID is a foreign key.
9:	If (ID.Found = 'True') Then
10:	Recuperate DFK () // Collect all data where ID is a Foreign Key From CSF.
11:	End
12:	Data Cleaning () // Correcting errors in the recovered data.
13:	Create Generic Classes () // Categorize the collected data.
14:	Search (ID, ADB) // Searching in ADB where ID is a Foreign Key.
15:	If (ID.Found = 'True') Then
16:	Recuperate ADB () // Recuperate data from All Data Bases.
17:	Data Cleaning ()
18:	Create R Classes () // Categorize related data.
19:	End
20:	Formulate C R () // Formulate Concepts and Relations.
21:	Formalize C () // Define concepts with a formal language.
22:	Integrate C () // Insert concepts in the ontology.
23:	ID = ID + 1
24:	End
25:	End

The data mining algorithm is comprised of three steps: public data collection, data classification, and knowledge integration, as described in Fig. 2. This proposed algorithm is a complete process in that it builds the national ontology dynamically from publicly available data sources.

3.1.1. Step 1: Public Data Collection

In this step, the data mining algorithm connects to publicly available data sources (such as civil status files, government databases, and so on) in order to dynamically collect (i.e. automatically and continuously) all data pertaining to citizens. After uploading the data sources, the algorithm starts by successively locating citizen national identification number (ID) in the civil status files in

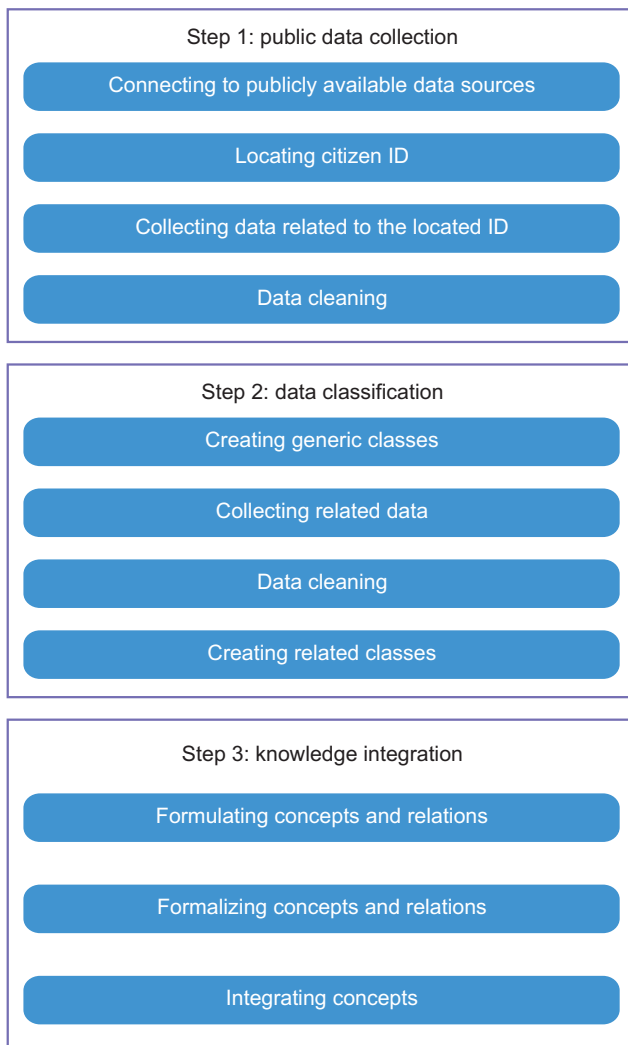


Fig. 2. The steps of the data mining algorithm. ID, national identification number.

order to retrieve all information (i.e. name, date of birth, place of birth, nationality, etc.) related to the selected ID as well as all ID information in relation with the located ID. The algorithm then searches for the citizen ID in all data sources (e.g., social security, financial institution, tax administration, etc.) in order to collect all data related to the located citizen ID. Once all data has been recovered, the data mining algorithm performs data cleaning, which consists of detecting and correcting (or removing) errors in the recovered data. This process involves the following sub-steps:

- Removing duplicate or irrelevant data: While we combine data sets from multiple sources, we may encounter duplicate data. Therefore, the algorithm removes unwanted findings from the recovered dataset, such as duplicate or irrelevant data.
- Fixing structural errors: Unusual naming conventions, typos, and incorrect capitalization will be corrected to avoid mislabeled categories or classes.
- Handling missing data: Missing values must not be ignored. Therefore, the algorithm will insert missing values based on other data. For example: when the first name value is missing, we insert it from another source where the ID is similar.

At the end of the data cleaning process, we will acquire coherent collected data (Fig. 3).

The public data collection is an ongoing process that involves collecting continuously and successively all data pertaining to citizens. Then, in the next algorithm step, the collected data will be classified and categorized automatically.

3.1.2. Step 2: Data Classification

This step aims to categorize the collected data using a data mining classification technique. First, after retrieving all information related to the selected citizen ID, the data mining algorithm creates generic classes that represent the data source from which we retrieve information. First and foremost, since we have acquired the data pertaining to the citizen ID from the civil status file, the algorithm will create the generic class 'Citizens', and then creates a sub-class of 'Citizens' with the retrieved information (i.e. name, date of birth, place of birth, nationality, etc.) as attributes. In addition, more 'Citizens' sub-classes of related IDs (i.e. foreign keys) will be created and relations between all sub-classes will be generated, as described in Fig. 4.

Moreover, and based on the searching results of the

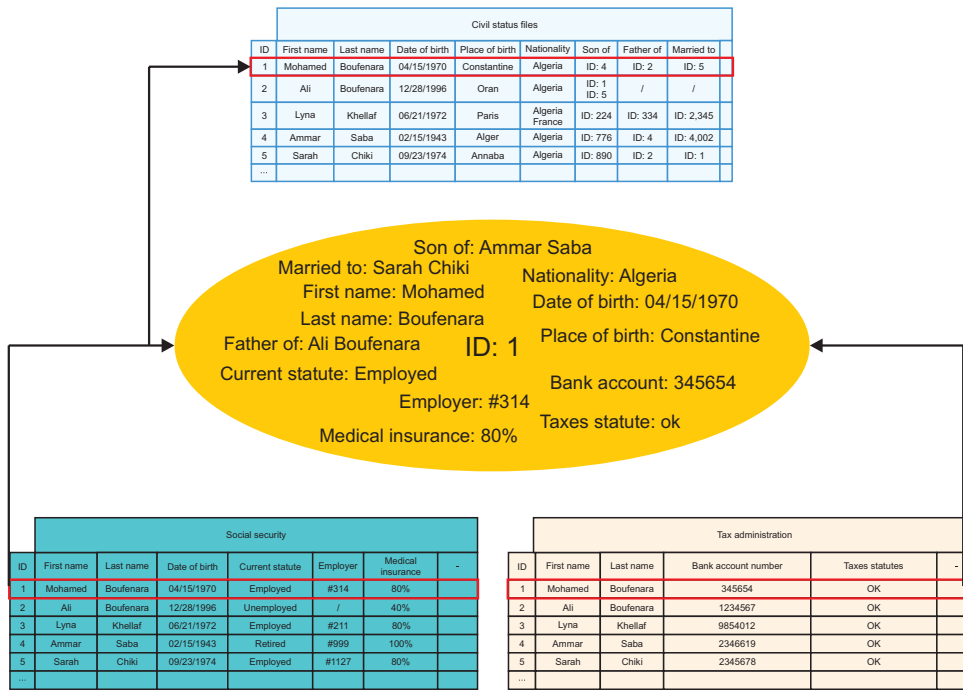


Fig. 3. The Public data collection. ID, national identification number.

citizen ID in all data sources (e.g., social security, financial institution, tax administration, etc.), the data mining algorithm will create classes of related information (Fig. 5).

Once all data has been classified and categorized, the data mining algorithm proceeds to the final step, which is to integrate the categorized knowledge, allowing the dynamic development of the national ontology.

3.1.3. Step 3: Knowledge Integration

The main task of this step is to automatically build the national ontology by structuring and integrating the retrieved data. The collected information is first formulated as ontological concepts by the Data Mining algorithm, with each class created in Step 2 converted into a concept and each correlation between classes converted into a relation. Then, using the syntax of the SHIQ description logic (Baader et al., 2017), the data mining algorithm formalizes concepts with a formal and operational language as described below:

1) Concept representation

ID1:

$(\forall \text{ Part of.Citizens}) \cap (\geq 1 \text{ Married to.ID5}) \cap \leq 1 \text{ Married to.ID5}) \cap (\exists \text{ Son of.ID4}) \cap (\exists \text{ Father of.ID2}) \cap (\geq 1 \text{ Works in.314}) \cap \leq 1 \text{ Works in.314}) \cap (\geq 1 \text{ have an account'.Bank1}) \cap \leq 1 \text{ have an account'.Bank1}) \cap (\geq 1 \text{ Pay taxes.Tax administration}) \cap \leq 1 \text{ Pay taxes.Tax administration}) \cap (\geq 1 \text{ Have insurance coverage.Social security}) \cap \leq 1 \text{ Have insurance coverage.Social security})$

1) Have insurance coverage .Social security)

2) Relation representation

Married to (ID1, ID5)

Son of (ID1, ID4)

Father of (ID1, ID2)

Works in (ID1, 314)

Have an account (ID1, Bank1)

Pay taxes (ID1, Tax administration)

Have insurance coverage (ID1, Social security)

Following concept formalization, we use the classification algorithm described below to perform concept integration:

Classification algorithm
1: Input: X // The concept to integrate.
2: Begin
3: Sps := SPS(X) ; // Search the super concepts of X
4: Spg := SPG(X) ; // Search the sub concepts of X
5: If ((Sps \cap Spg) = \emptyset) Then // X is a new concept to integrate.
6: For each (S \in Sps) Do
7: Add (X \subseteq S) ; // Adding the subsumption link X \subseteq S.
8: End
9: For each (G \in Spg) Do
10: Add (G \subseteq X) ; // Adding the subsumption link G \subseteq X.
11: End
12: For each (S \in Sps) Do
13: For each (G \in Spg) Do
14: Remove (G \subseteq S) ; // Removing the subsumption link G \subseteq S.
15: End
16: End
17: Else // X is an existing concept.
18: Return (nil);
19: End if
20: End

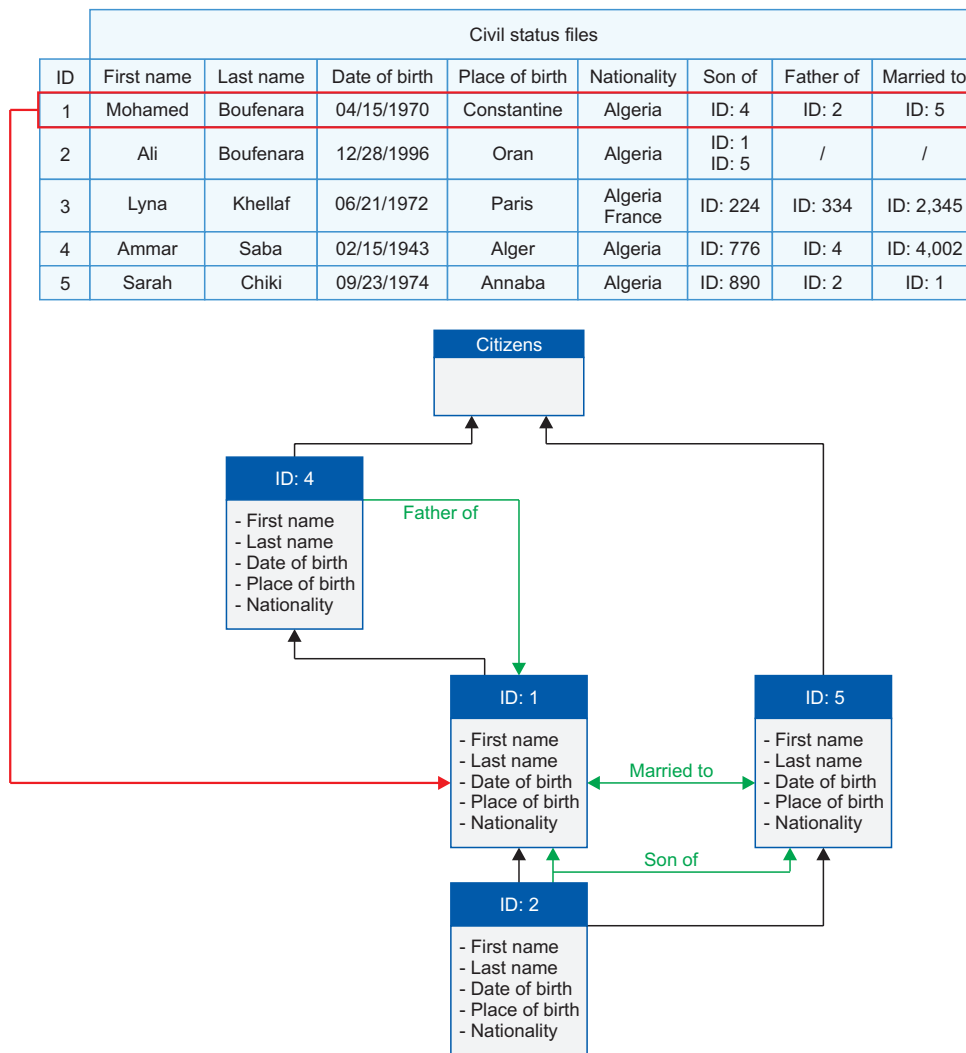


Fig. 4. Citizen classes creation.
ID: national identification number.

The knowledge integration process will enable the ontology to be constructed automatically and continuously. The developed ontology and its features will be presented in the following section.

3.2. The National Ontology

An ontology is, in its most basic form, a table of categories in which each type of entity is represented by a node in a hierarchical tree. We rely on the national ontology developed by the data mining algorithm to achieve our goal of providing authorities with quantitative and qualitative data about any country's social and economic situation. The national ontology serves as a reference knowledge base that can be used to provide reliable statistical knowledge to governmental, economic, and social actors. It includes concepts (type definitions) that are data descriptors for citizen information, as well as the relation-

ships between these concepts. The following are the key features of our national ontology:

- A hierarchical classification of concepts (classes) from general to specific.
- A list of attributes for each class.
- A set of relations between classes to link concepts.

In the national ontology, various generic classes serve to define intricate concepts like 'citizens,' 'employer,' 'social security,' 'banks,' and 'tax administration.' These classes are combined using data correlation, as illustrated in Fig. 6.

The ontology tree will be restructured based on the classification algorithm presented above each time the data mining algorithm collects, categorizes, and integrates concepts. New concepts will be structured and integrated into the ontology as subclasses of the generic classes.

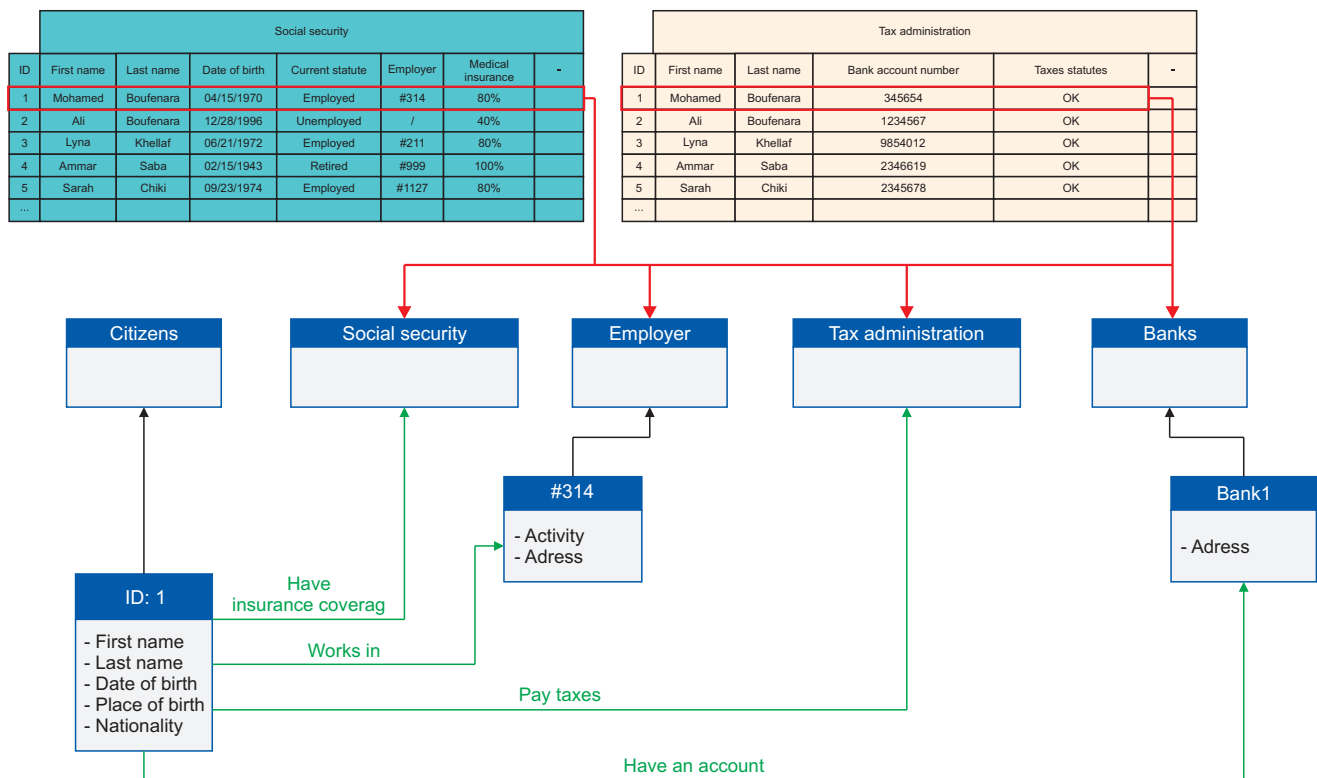


Fig. 5. Classes structuring. ID, national identification number.

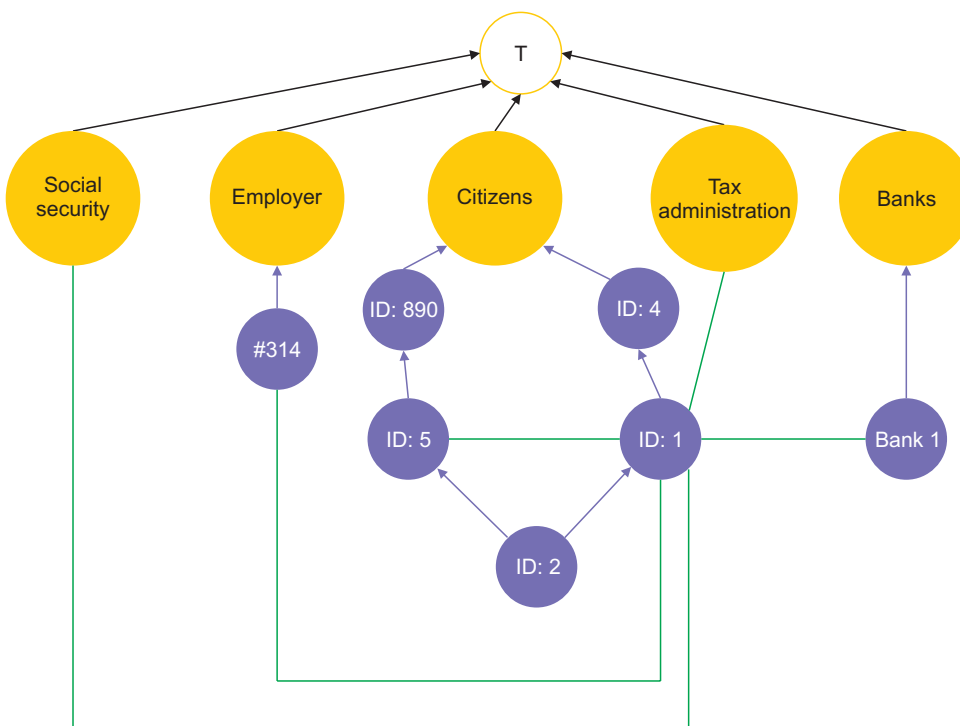


Fig. 6. The ontology concepts tree. ID, national identification number.

The national ontology is being created dynamically (i.e. automatically and continuously) and will be used as a reference knowledge base to provide reliable statistical knowledge to governmental, economic, and social actors. Therefore, we propose an intelligent exploitation platform to enable the use of the developed national ontology.

3.3. The Intelligent Ontology Exploitation Platform

Our proposed MVC architecture presents the view as an ontology exploitation platform that utilizes artificial intelligence for autonomy, information exchange, and co-operative management. The intelligent platform provides two features that allow users to interact with the national ontology. The first feature enables citizens and government administrators to explore and consult available knowledge, facilitating the digitization of administrative tasks and citizens' daily lives. The second feature continuously offers quantitative and qualitative information about the country's socioeconomic situation to governmental, economic, and social actors, as illustrated in Fig. 7.

Access to the platform is granted to government administrators based on their profile. For example, police services can only access legal, criminal, and civil status information about citizens. On the other hand, tax services have access to all information pertaining to citizens' earnings. In addition, governmental, economic, and social actors have privileged access to the statistical information provided by the platform. These quantitative and qualitative knowledge assets about the country's socioeconomic situation are continuously produced, analyzed, and provided by a platform module interacting with the national

ontology and applying both quantitative and qualitative statistical techniques.

Since the national ontology consists of concepts that serve as data descriptors for citizen information, as well as the relationships between these concepts, the platform module will consider the national ontology as the population and will perform quantitative and qualitative analysis on it. To begin, the platform module utilizes different types of variables, including discrete and continuous quantitative variables and nominal and ordinal qualitative variables. These variables are detailed in Table 1.

The platform module can also use multiple variables at the same time. These variables can be qualitative, quantitative, or a mix of both. For example, the module combines multiple statistical variables to determine the number of citizens earning at least the minimum wage, the number of unemployed women, the number of low-income citizens, and so on. To acquire such statistical knowledge, the platform module constantly queries the national ontology using the SPARQL querying language, which is a standardized language that is similar to SQL but is used for Resource Description Format (RDF) graphs.

The national ontology is an OWL ontology in RDF that can be queried as an RDF graph using the SPARQL query language. SPARQL provides a means to query diverse data sources by allowing the expression of both mandatory and optional graph patterns, as well as their combinations. Additionally, it includes aggregate functions that can be used to select and return calculated values based on groups of query results. SPARQL can generate statistical data by using common aggregate functions such as: COUNT, SUM,

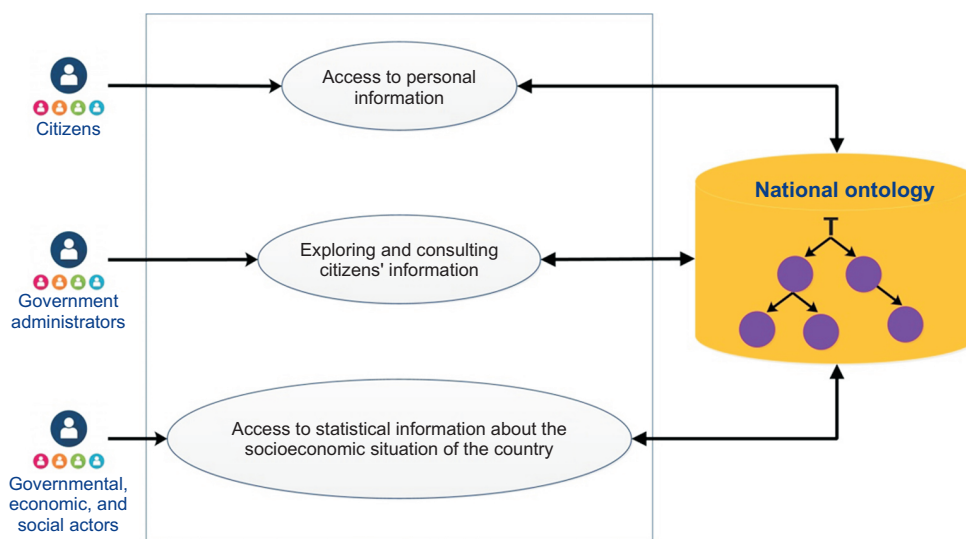


Fig. 7. Use cases of the intelligent platform.

Table 1. Quantitative and qualitative variables

Quantitative		Qualitative	
Discrete	Continuous	Nominal	Ordinal
All numerical values that can be counted	All numerical values that can be measured	Labels that describe groups of observations without logical order	Labels that describe groups of observations with a logical order
Examples: A country's total population A city's total population Total number of voters Total number of citizens with medical coverage	Examples: Citizens' payroll Citizens' tax amount Tax evasion amount	Examples: Employed citizens Unemployed citizens Married, unmarried	Examples: Level of tax evasion (high>medium>low) Level of citizens without medical coverage (high>medium>low)

MIN, MAX, AVG, DISTINCT, HAVING, and GROUP BY. The platform module employs these aggregate functions to generate reliable statistical knowledge, as illustrated in the following examples:

- 1) To determine a country's total population, the platform constantly queries the national ontology by counting instances belonging to the "Citizens" class.

```
SELECT DISTINCT ?Citizens (COUNT(?Citizens) as ?total population)
```

- 2) To determine the total population of a city, the national ontology will be queried by counting instances of the "Citizens" class with the given city as an attribute.

```
SELECT DISTINCT ?Citizens (COUNT(?Citizens) as ?CTP)
WHERE {
    ?citizens:city 'Paris' .
```

- 3) The platform module queries the national ontology for the number of unemployed women by counting instances of the "Citizens" class with female sex and unemployed status as attributes.

```
SELECT DISTINCT ?Citizens (COUNT(?Citizens) as ?UW)
WHERE {
    ?citizens:sex 'female' ?citizens:status 'unemployed.'
```

Following the SPARQL querying of the national ontology, graphical representations will be charted to provide reliable and understandable statistical knowledge, allowing for better management and decision making.

4. SOFTWARE APPLICATION AND EXPERIMENT

We developed a software application to simulate and to

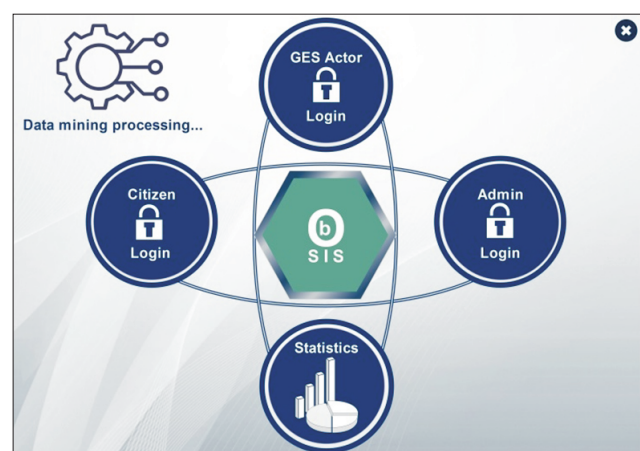


Fig. 8. The system interface. GES, governmental, economic, and social actors; OBSIS, ontology-based statistical information system.

experiment with our proposed data mining approach for the dynamic development of an ontology-based statistical information system. We performed a case study on Algerian citizens, using publicly Algerian data sources (i.e. civil status files, government databases, etc.).

The software application was designed as a back and front interface (Fig. 8) that enables the dynamic ontology development as well as the use of the national ontology.

The interface back-end represents the behavior of the data mining algorithm, which automatically and continuously collects, classifies, and categorizes citizens' knowledge into the ontology, as described in Fig. 9.

The proposed intelligent ontology exploitation platform (Fig. 10) serves as the interface's front end, enabling citizens, government administrators, and governmental, economic, and social actors to use the national ontology.

Citizens can log in to their personal profiles to manage their daily paperwork, applications for employment, hous-

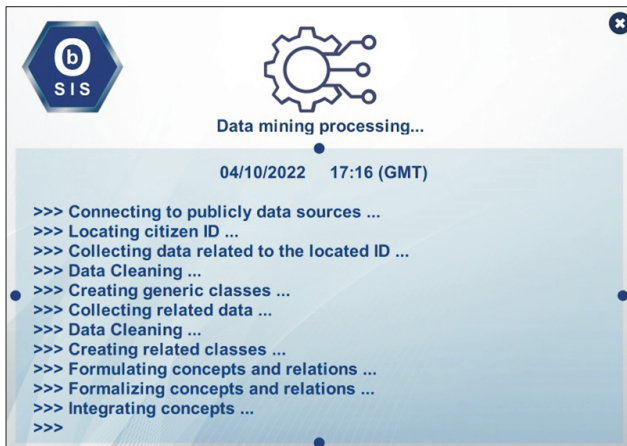


Fig. 9. The interface back-end. GMT, greenwich mean time; ID, national identification number.



Fig. 11. Citizen profile. OBSIS, ontology-based statistical information system; ID, national identification number.



Fig. 10. The interface front-end. GES, governmental, economic, and social actors; OBSIS, ontology-based statistical information system.



Fig. 12. Admin access. OBSIS, ontology-based statistical information system; ID, national identification number.

ing, and assistance, and so on (Fig. 11).

Moreover, administrators representing governmental services such as health and social care, justice, taxes, police, and so on can interact with “the national ontology” by exploring or consulting available knowledge, which will aid in the digitization of administrative and governmental tasks. Fig. 12 depicts a scenario in which the police services use the platform to obtain information about a citizen.

In addition to and as illustrated in Fig. 13, governmental, economic, and social actors have privileged access to the platform’s statistical information on the country’s socioeconomic situation.

To experiment with our proposed approach, we performed a case study on Algerian citizens, using publicly

Algerian data sources in order to constantly develop a national ontology. Statistical knowledge assets about Algeria’s socioeconomic situation are continuously produced, analyzed, and provided by SPARQL querying this ontology and applying both quantitative and qualitative statistical variables. Besides this, graphical representations will be charted to provide reliable and understandable statistical knowledge, allowing for better management and decision making, as described below:

- The average amount of unemployed citizens.

```
SELECT (AVG(?UC) AS ?avg-Unemployed )
WHERE {
  SELECT DISTINCT ?Citizens (COUNT(?Citizens) as ?UC)
  WHERE { ?citizens:status 'unemployed' }
}
```

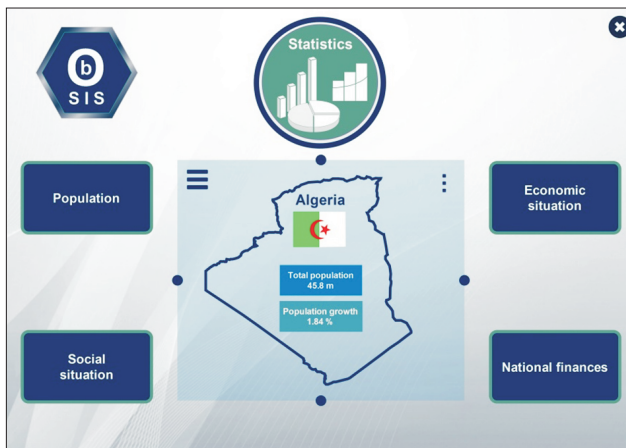


Fig. 13. Access to statistics. OBSIS, ontology-based statistical information system.

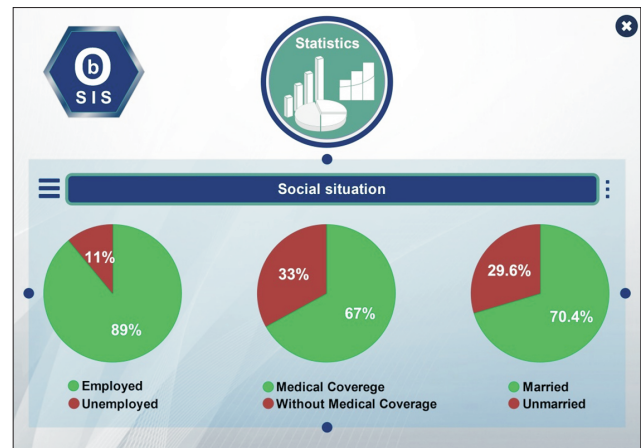


Fig. 14. Social situation parameters. OBSIS, ontology-based statistical information system.

- The average amount of unmarried citizens.

```
SELECT (AVG(?MW) AS ?avg-Min-wage )
WHERE {
  SELECT DISTINCT ?Citizens (COUNT(?Citizens) as ?MW)
  WHERE {
    ?citizens:Marital-status 'unmarried' }
}
```

- The average amount of citizens without medical coverage.

```
SELECT (AVG(?WMI) AS ?avg-Without-M-Insurance )
WHERE {
  SELECT DISTINCT ?Citizens (COUNT(?Citizens) as ?WMI)
  WHERE { FILTER (?medical insurance = 0) }
}
```

This ontology querying allows the platform to display graphical representations of some social situation parameters, as shown in Fig. 14.

Furthermore, the following SPARQL queries can be used to obtain some economic situation parameters:

- The proportion of citizens earning the minimum wage.

```
SELECT (AVG(?MW) AS ?avg-Min-wage )
WHERE {
  SELECT DISTINCT ?Citizens (COUNT(?Citizens) as ?MW)
  WHERE { FILTER (?salary = 20000) } }
```

- The percentage of citizens earning the average salary.

```
SELECT (AVG(?AS) AS ?avg-med-wage )
WHERE {
  SELECT DISTINCT ?Citizens (COUNT(?Citizens) as ?AS)
  WHERE { FILTER (AVG(?salary) ) } }
```

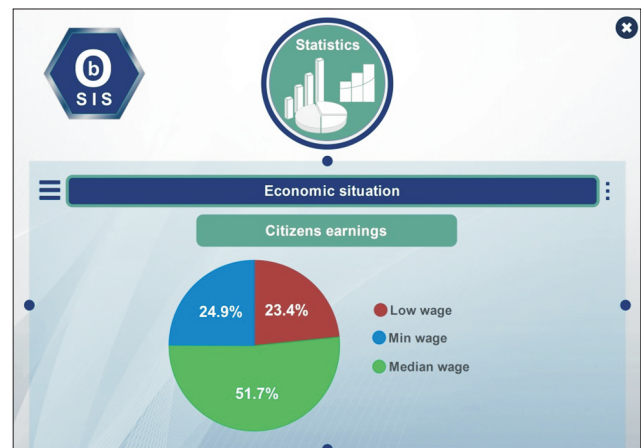


Fig. 15. Economic situation parameters. OBSIS, ontology-based statistical information system.

- The average amount of low-income citizens.

```
SELECT (AVG(?LIC) AS ?low-income )
WHERE {
  SELECT DISTINCT ?Citizens (COUNT(?Citizens) as ?LIC)
  WHERE { FILTER (?salary < 15000) } }
```

Based on the outcomes of these queries, graphical representations of the economic situation will be charted, as shown in Fig. 15.

More SPARQL queries will be applied in order to constantly obtain reliable quantitative and qualitative knowledge about the country's socioeconomic situation. This produced statistical knowledge will be used by governmental, economic, and social actors for strategic decision-making, administration, and management of the national

political, social, and economic life.

5. DISCUSSION

Statistical data are used in many sectors, including science, government, manufacturing, population, psychology, banking, and financial markets. Statistics is the science that allows data to be identified, collected, organized, interpreted, and presented. Data can be either qualitative or quantitative. Statistics facilitate information-based decision-making. Indeed, since they collect, store, transform, and provide statistical data, both statistical Information Systems and Data Mining methods have played important roles in improving decision-making. Several statistical information systems and statistical data mining approaches have been developed as a result of these challenges. In Table 2, we compare some of these approaches by Bacchi et al. (2022), Improt et al. (2021), Ngugi (2021), Ngugi et al. (2021), and Osi et al. (2020) with our own method.

Our proposed approach employs a data mining technique to dynamically develop an ontology-based statistical information system. In comparison to some existing solutions that develop static statistical information systems, our developed system allows for the automatic and continuous collection, storage, processing, analysis, and presentation of statistical data at the national scale. To

reach this goal, a classification data mining technique was developed to automatically collect, classify, and integrate citizen data into a national ontology. Then, an intelligent platform was developed to generate quantitative and qualitative statistical information based on the knowledge stored in the ontology in order to provide instant reliable knowledge about the country's socioeconomic situation. The provided statistical information will be used by governmental, economic, and social actors for strategic decision-making, administration, and management of national political, social, and economic life. In addition, as illustrated by the software application and experiment, the proposed approach has successfully digitized administrative and governmental tasks, thereby making citizens' lives easier. Moreover, the proposed system has successfully provided reliable and understandable statistical knowledge with graphical representations, allowing for better management and decision making.

6. CONCLUSION

In the age of big data, statistical knowledge is a valuable resource in decision-making. Statistics are becoming increasingly important in decision-making as an increasing amount of data is collected and converted into usable information and actionable steps. As a result, recent

Table 2. A Comparison of Some Existing Approaches

Approaches	Statistical data mining approaches			Developing statistical information systems	
	Classification	Clustering	Prediction	Static	Dynamic
Adekitan and Noma-Osaghae, 2019	X	X	X		
Luković et al., 2019				X	
Baek et al., 2018	X		X		
Ledvinka et al., 2019				X	
Fernandes et al., 2019			X		
Comas Rodríguez et al., 2019				X	
Jatav, 2018			X		
Choi and Choi, 2019				X	
Improt et al., 2021			X		
Ngugi, 2021				X	
Bacchi et al., 2022			X		
Ngugi et al., 2021				X	
Osi et al., 2020	X		X		
Our approach	X				X

research has focused on using computational techniques from statistics, data mining, and information theory to automatically collect and analyze massive amounts of data, thereby improving the availability of information on the state and evolution of a social or economic situation. Our proposed data mining approach for the dynamic development of an ontology-based statistical information system is part of these research efforts to provide governmental, economic, and social authorities with reliable statistical information in order to improve management and decision making. Unlike some existing statistical information solutions, the proposed ontology-based statistical information system enabled the national scale automatic and continuous collection, storage, processing, analysis, and presentation of statistical data. We proposed an MVC-inspired approach for developing our system, with the Controller representing a data mining algorithm that dynamically (i.e. automatically and continuously) collects all data pertaining to citizens from publicly available data sources (i.e. civil status files, government databases, etc.), and then structures, classifies, and categorizes this knowledge into the ontology. The Model is a national ontology that incorporates all citizen knowledge, and the View is an intelligent ontology exploitation platform that provides authorities with quantitative and qualitative data about the country's social and economic situation. To evaluate the proposed approach, a case study on Algerian citizens was conducted, using publicly available Algerian data sources to continuously develop a national ontology. SPARQL queried the developed ontology and applied both quantitative and qualitative statistical variables to continuously produce, analyze, and provide statistical knowledge about Algeria's socioeconomic situation. Furthermore, in addition to demonstrating the feasibility of our approach, the software application with the modeled system have demonstrated additional benefits such as the digitization of administrative and governmental tasks, thereby making citizens' lives easier, and providing reliable and understandable statistical knowledge with graphical representations, allowing governmental, economic, and social actors to better manage and make decisions.

ACKNOWLEDGEMENTS

The authors acknowledge support from the General Directorate of Scientific Research and Technological Development (DGRSDT), Ministry of Higher Education and Scientific Research, Algeria.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

- Adekitan, A. I., & Noma-Osaghae, E. (2019). Data mining approach to predicting the performance of first year student in a university using the admission requirements. *Education and Information Technologies*, 24(2), 1527-1543. <https://doi.org/10.1007/s10639-018-9839-7>
- Baader, F., Horrocks, I., Lutz, C., & Sattler, U. (2017). *An introduction to description logic*. Cambridge University Press.
- Bacchi, S., Tan, Y., Oakden-Rayner, L., Jannes, J., Kleinig, T., & Koblar, S. (2022). Machine learning in the prediction of medical inpatient length of stay. *Internal Medicine Journal*, 52(2), 176-185. <https://doi.org/10.1111/imj.14962>
- Baek, H., Cho, M., Kim, S., Hwang, H., Song, M., & Yoo, S. (2018). Analysis of length of hospital stay using electronic health records: A statistical and data mining approach. *PLoS One*, 13(4), e0195901. <https://doi.org/10.1371/journal.pone.0195901>
- Choi, C., & Choi, J. (2019). Ontology-based security context reasoning for power IoT-cloud security service. *IEEE Access*, 7, 110510-110517. <https://doi.org/10.1109/ACCESS.2019.2933859>
- Comas Rodríguez, R., Simón-Cuevas, A., Sánchez Fleitas, N., & García Lorenzo, M. M. (2019, October 28-31). An ontology-based data management model applied to a real information system. In J. Nummenmaa, F. Pérez-González, B. Domenech-Lega, J. Vaunat, & F. O. Fernández-Peña (Eds.), *Proceedings of the Conference on Computer Science, Electronics and Industrial Engineering (CSEI 2019)* (pp. 36-50). Springer.
- Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research*, 94, 335-343. <https://doi.org/10.1016/j.jbusres.2018.02.012>
- Improta, G., Colella, Y., Rossi, G., Borrelli, A., Russo, G., & Triassi, M. (2021, October 29-31). Use of machine learning to predict abandonment rates in an emergency department. *Proceedings of the 2021 10th International Conference on Bioinformatics and Biomedical Science (ICBBS '21)* (pp. 153-156). ACM.
- Jatav, S. (2018). An algorithm for predictive data mining approach in medical diagnosis. *International Journal of Com-*

- puter Science & Information Technology, 10(1). <https://ssrn.com/abstract=3633801>
- Kermani, M. H., & Boufaïda, Z. (2022). I 3D3P: An intelligent 3D protein prediction platform. *Proceedings of International Conference on Applied Innovation in IT*, 10(1), 37-42. <https://doi.org/10.25673/76930>
- Kermani, M. H., Boufaïda, Z., Benredjem, S., & Saker, A. N. (2021). An MVC-inspired approach for an intelligent annotation of a protein ontology: IA-PrOnto. *International Journal of Computer Information Systems and Industrial Management Applications*, 13, 308-318. http://www.mir-labs.org/ijcism/regular_papers_2021/IJCISIM_29.pdf
- Ledvinka, M., Lališ, A., & Křemen, P. (2019). Toward data-driven safety: An ontology-based information system. *Journal of Aerospace Information Systems*, 16(1), 22-36. <https://doi.org/10.2514/1.I010622>
- Luković, V., Ćuković, S., Milošević, D., & Devedžić, G. (2019). An ontology-based module of the information system ScolioMedIS for 3D digital diagnosis of adolescent scoliosis. *Computer Methods and Programs in Biomedicine*, 178, 247-263. <https://doi.org/10.1016/j.cmpb.2019.06.027>
- Nawi, R. M., Noah, S. A. M., & Zakaria, L. Q. (2021). Issues and challenges in the extraction and mapping of linked open data resources with recommender systems datasets. *Journal of Information Science Theory and Practice*, 9(2), 66-82. <https://doi.org/10.1633/JISTaP.2021.9.2.5>
- Ngugi, P., Babic, A., & Were, M. C. (2021). A multivariate statistical evaluation of actual use of electronic health record systems implementations in Kenya. *PLoS One*, 16(9), e0256799. <https://doi.org/10.1371/journal.pone.0256799>
- Ngugi, P. N. (2021). *A systematic method for evaluating implementations of electronic medical records systems in low- and medium-income countries*. (Doctoral dissertation). University of Bergen, Bergen, Norway.
- Osi, A. A., Abdu, M., Muhammad, U., Ibrahim, A., Ismail, L. A., Suleiman, A. A., Abdulkadir, H. S., Sada, S. S., Dikko, H. G., & Ringim, M. Z. (2020). A classification approach for predicting COVID-19 patient's survival outcome with machine learning techniques. *medRxiv*. <https://doi.org/10.1101/2020.08.02.20129767>
- Plirdpring, P., & Ruangrajitpakorn, T. (2022). Using ontology to represent cultural aspects of local products for supporting local community enterprise in Thailand. *Journal of Information Science Theory and Practice*, 10(1), 45-58. <https://doi.org/10.1633/JISTaP.2022.10.1.4>
- Sowa, J. F. (2011). *Serious semantics, serious ontologies, panel*. Paper presented at the Semantic Technology Conference (SEMTEC 2011), San Francisco, CA, USA.