Knowledge-based Big Data Cleanup Method

Andrii Berko[0000-0001-6756-5661]1 , Vladyslav Alieksieiev [0000-0003-0712-0120]2 ,

Vasyl Lytvyn [0000-0002-9676-0180]3

Lviv Polytechnic National University, Lviv, Ukraine
andrii.y.berko@lpnu.ua[1],
vladyslav.i.alieksieiev@lpnu.ua[2]
Vasyl.V.Lytvyn@lpnu.ua[3]

Abstract. Unlike traditional databases, Big Data stored as NoSQL data resources. Therefore such resources are not ready for efficient use in its original form in most cases. It is due to the availability of various kinds of data anomalies. Most of these anomalies are such as data duplication, ambiguity, inaccuracy, contradiction, absence, the incompleteness of data, etc. To eliminate such incorrectness, data source special cleanup procedures are needed. Data cleanup process requires additional information about the composition, content, meaning, and function of this Big Data resource. Using the special knowledge base can provide a resolving of such problem.

Keywords: Big Data, Ontology, Knowledge Base, Data Cleanup.

1 Introduction

The problem of data quality remains to be topically for a long time in various areas of data processing. Nowadays it is especially considerable in Big Data technologies and analytics [5,6]. The particularity of the problem of data quality is researched and discussed in [1,2,4,10]. Not only volume, variety, and velocity of changes are the principal quality factors for information resources developed on the principles of Big Data. The syntax and content heterogeneity of the resources themselves, the complexity of control, influence and management of the processes of their production and development also take place [2]. These factors, often contribute to the emergence of some data item corruptions [1,2] in the information resource content. Generally, Big Data resources are presented in NoSQL database formats. It means that principal requirements for such data resources are availability and partition tolerance. At the same time, any consistency constraints are not supported for such data [6,11]. No any constraints such as check action, unique, and not null requirements are not used in NoSQL data resources too. Weak data consistency leads to a situation when data resources are not ready for use in the original form. So, Big Data resources obtained in NoSQL formats are not checked, refined, and consistent so fine as traditional databases. The consequence of this is the risk of occurrence incorrect, inconsistent or invalid data values in a data set[4,6,9]. Therefore some steps to prepare these resource for efficient processing are needed. It means some data values must be transformed to the form corresponded with data source purpose, meaning of the tasks, and user requirements during preparing processes. One of the principal steps of Big Data resource preparing for its use is the application of data cleanup actions. Incorrect or invalid data values have to be edited, corrected or replaced by right and valid values at the clean-up stage of Big Data resource [5,9]. As a result, we can obtain the set of so-called "clean" data, which are correct, valid and ready to use according to their functions [1,4]. One of the principal problems of Big Data cleanup process is the details formal description of all data properties, features of data items invalidity, and efficient cleanup actions.

2 Data Anomalies Processing in Big Data Resources

Data anomalies in the Big Data resources are presented by such phenomena as absence, duplication, ambiguity, lack of meaning, inaccuracy, incompleteness, unreliability, inconsistency, etc. [3,9]. The existence of data abnormalities greatly degrades the consumer properties of information resources, makes it difficult or impossible to efficient using due to invalid data items presence. The consequence of this is the incorrect execution of operations for the search, selection or analysis of data. For example, we would process values that are equal to each other as different due to their inaccuracy, misrepresentation, corruption or input error, when performing such operation as data mapping – Map(X) of the MapReduce [5,6] method. For the same reasons, different values would be mistakenly presented as equal. The absence or inadmissibility of some values makes it impossible to use them, etc. Therefore, the correct and efficient work with the Big Data provides for the procedures of their cleanup, during which, in particular, perform the elimination of existing data anomalies. Data anomalies interpretation is one of the principal tasks for efficient Big Data resource cleanup. The interpretation of data anomalies depends on their nature and the causes of the occurrence. It allows to recognize data anomalies in Big Data resource correctly and to choose the most suitable method of this anomalies elimination. For a successful solution of data anomalies problem, these need to be classified. According to [1,3] such principal types of data anomalies are defined for big data resources::

- value not present,
- value is unknown,
- value is invalid,
- value is duplicated,
- value is ambiguous,
- value is not accurate enough,
- value is an incomplete,
- value is unreliable etc.

Classification allows us to choose the best way to eliminate it of data anomalies.

Each data anomaly would be detected for a data item by checking its correspondence to some predefined requirements. Such requirements have to be described as conditions of data value comparison. For detection of described above types of anomalies, such conditions may be used [1] (Table 1).

Table 1. Correspondence between data anomalies and check conditions

Type of data anomalies	Data item check condition
value not present value is unknown	data item Is Null
value is invalid	data item Is not in <interval>
value is unreliable	data item Is not in <set>
value is duplicated	Count(data item)>1
value is ambiguous	data item != data item
value is not accurate enough	data item != value
value is an incomplete	Number(data item)<value

This list may be continued or changed according to specifics of processed data.

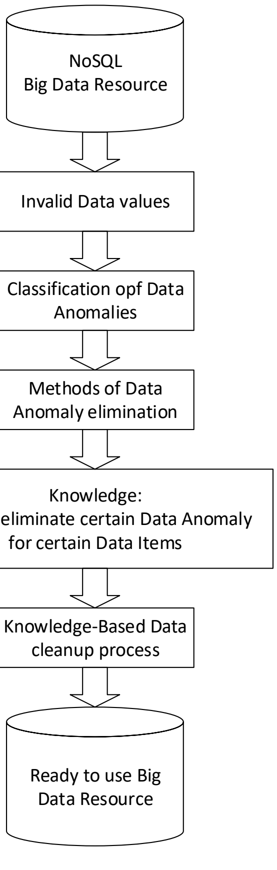

NoSQL
Big Data Resource
Invalid Data values
Classification opf Data Anomalies
Methods of Data Anomaly elimination
Knowledge:
How to eliminate certain Data Anomaly for certain Data Items
Knowledge-Based Data cleanup process
Ready to use Big Data Resource

Fig. 1. General schema of producing and use of Big Data resource cleanup knowledge

Usually, to eliminate data anomalies, the most commonly used techniques are the removal, ignoring or re-defining of an appropriate data element, using the average, most likely, estimated or surrogate value, duplicate values remove etc. [1,4,9].

The principal problem of efficient Big Data resource cleanup is how to build exact and complete descriptions of the methods and rules of elimination of data anomalies. This may be presented as a specific knowledge set includes:

1. a description of Big Data resource, its properties, and all included data items,
2. description of data anomaly types,
3. rules of search and recognition of invalid data values,
4. description of the methods of fixing corrupted or invalid data.

General schema of producing and use of such knowledge presented on Fig. 1.

The process of anomalies of data elimination in the Big Data resource performs as a replacement of the incorrect value by the new value, which define by a special procedure. In the general case, the value v_{ij} of some data unit V_i, which is formed to eliminate its anomaly, depends on the nature (category) of data anomalies – U_k and the method of its elimination – S_l. The procedure for defining a new data value can be describe as a sequence of steps of kind

$$V_i \;\rightarrow\; U_k \rightarrow S_l \rightarrow v_i\,. \tag{1}$$

That mean: invalid data value V_i of category U_k by using of method S_l have to be replaced by value v_i for elimination of one case of data anomalies in some resource. The same transformation can be presented as a mapping

$$v_i = \Phi(V_i\,, U_k, S_l), \tag{2}$$

where Φ is a function for define new value for invalid data item using its category and corresponding method. These actions are perform during general data cleanup process of Big Data source. Using ontologies for Big Data cleanup

Because it is necessary to have exact and complete descriptions of the correspondence between anomalies in the data resource, their classification is needed. As well a formal description of the ways to eliminate such anomalies is needed. Knowledge base may be uses for such purpose in the set of tools for Big Data sources cleanup. The core of this knowledge base may be formed by an ontology of type

$$O^O =< C^O, R^O, F^O>, \tag{3}$$

where $C^O=\{C^V, C^U, C^S\}$ is the set of concepts (classes), which include such subclasses:
C^V is entity set (subclass) for a presentation of data units in the Big Data resource to be processed,
C^U is the set of entities that describe the types and nature of each of the data anomalies presented in the Big Data resource,
C^S is the set of entities for the description of data anomalies elimination methods (data anomaly problems solving);
$R^O = \{R^{VU}, R^{US}, R^{VUS}\}$ is a set of relations between the above-defined concepts that include three subsets:

R^{VU} is a subset of binary relations between data values of C^V and types of anomalies in Big Data resource C^U,
R^{US} is a subset of relation between types of data anomalies of C^U and methods of anomaly elimination C^S,
R^{VUS} is a subset of ternary relations between data values and methods of data anomaly problem solving;
$F^O = \{ F^V, F^C \}$ is a set of axioms (rules). Two kinds of rules are needed for data cleanup process. Each one describes the steps which must be performed for solving of data anomaly problem, The first kind of rules – F^V include rules for data validation in presented Big Data resource. The second subset – F^C consist of rules defines the method (from the set C^S) of data cleanup. This method directly depends on certain data values (presented by the set C^V) and a certain type of anomalies of this data item (described by the set C^U).

The resource of Big Data needs to be deeply investigated to solve the problem of data cleanup. The main goal of this investigation is to get the answer to the questions:

- what data items and values are needed to solve some defined task;
- what is the structure and content of the data source used for this purpose;
- what kinds of problems with data are possible in the presented resource;
- what factors have an influence on data quality in the resource;
- what are data quality criteria and requirements;
- what are the methods to fix corrupted or invalid data items;
- how some methods of bad data fixing would be used to correct some data items.

Whether obtained all necessary information about input Big Data resource, we can formalize such knowledge as units of certain special ontology [1].

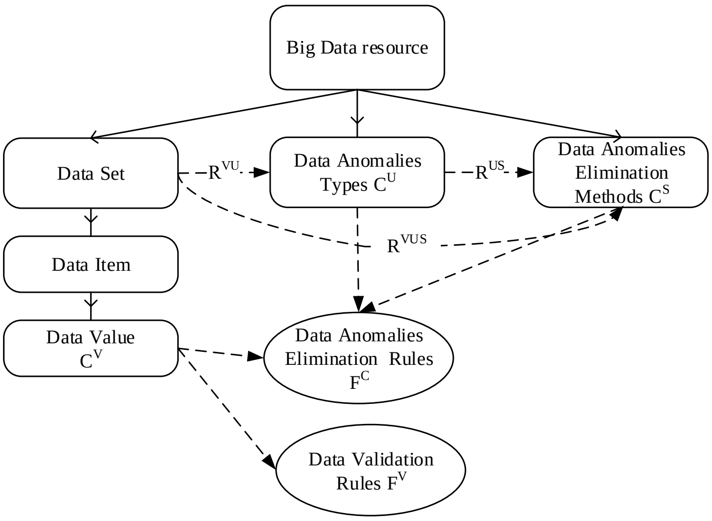

Big Data resource
Data Set
R^VU
Data Anomalies Types C^U
R^US
Data Anomalies Elimination Methods C^S
R^VUS
Data Item
Data Value C^V
Data Anomalies Elimination Rules F^C
Data Validation Rules F^V

Fig. 2. Structural model of ontology for Big Data resource cleanup

Generalized structural model of ontology for Big Data resources cleanup presented on Fig. 2. So, Big Data resource would be described by three concepts of the ontology according to the developed model. The first describes certain data as a hierarchy: "Data set" (table, collection, etc) –"Data item" (column, field, element) – "Data value". The second concept describes the types of potential anomalies of data for processed data resource. And the third one - the ways to fix corrupted or invalid data in the resource. Two types of axioms describe the rules of data check to find certain anomalies and the rules of various data anomalies elimination. The arches describe possible relations between defined concepts and the rules. As a remark, the relation R^{VUS} seems redundant, because of its matching to R^{VU} and R^{US}. Bet such overage allows defining of unambiguous correspondence between the data set and possible methods of data anomalies elimination.

According to proposed principles of the generalized model, the ontology for any Big Data resource is developed. It is necessary to determine the specific values of concepts, relationships, and rules in accordance with the content of the resource data when performing this action. As a result, the primary version of ontology for Big Data resource cleanup will be obtained. As the next, we can create an appropriate knowledge base for the data cleanup tools based on created ontology.

So, in the above-described way we can obtain the set of tools needed for efficient solving of Big Data resource anomalies elimination. Therefore, an ontology created in accordance with the principles described above can be used as the basis of the knowledge base for intelligent tools of the Big Data resources cleanup.

3 Algorithms and Tools for Big Data Resource Cleanup

The use of ontology as the core of the knowledge base of the Big Data Cleanup tools determines the peculiarities of the process of solving data problems (Fig. 3). So, construction and tuning of basic ontology is a principal prerequisite for any Big Data source cleanup process. This stage requires a number of actions based on expert knowledge. Expert knowledge provide definition and creation of basic ontology parts just like that.

1. Description of the set of concepts, which corresponds to data values and data units C^V, according to the set of requirements of ontology construction. The concept C^V for the presentation of data value may be defined as a result of hierarchic taxonomy – "*Data Set -> Data Item -> Data Value*" according to the proposed structural model of ontology.
2. Construction of the set of definitions of data anomalies C^U, which are characteristic of the given Big Data source. The list of most common data anomalies is presented above.
3. Definition of the set of methods C^S for data anomalies elimination. Most of well-known and often used are such methods of solution of data anomaly problem [1]:

- *repeat a request* to receive corrupted value,
- *recalculate* inaccurate value,
- *refine* inconsistent value,

- *replace* absent value with some aggregate value (average value, probable value, standard or default value, initial value, some calculated value, estimated value, expert value, etc.),
- *use* of artificial *surrogate marks* instead of absent or corrupted value,
- *remove* of corrupted/duplicated data item from the resource,
- *ignore* the data anomaly for given data item,
- *using* of *special tools* to process uncertainties

4. Definition of relations between concepts - data items, data anomalies and methods of data anomalies elimination – R^{VU}, R^{US}, R^{VUS}. The most suitable format for definition relations between concepts is RDF triplet [8] "*Object - Predicate - Subject*". So, these relations may be formed in such a way.

- For relation R^{VU} the triplets may be constructed by the scheme "*Data Value* (C^V)– *Have the Anomaly – Anomaly Type* (C^U)".
- For relation R^{US} – by the scheme "*Anomaly Type*(C^U) – *Eliminated by – Method* (C^S)".
- For relation R^{VUS} the triplets have to be constructed by the scheme "*Data Value* (C^V)– *Anomaly Type* (C^U) – *Method* (C^S)". Here, the concept C^U execute a function of the predicate.

5. Rules of data validation for its anomaly detection – F^V. This rules also may be presented as RDF triplets [8] by the scheme "*Condition – Corresponds to – Anomaly Type*". For example, the set of such rules may be like as the next

- *Value is null,* Corresponds to, *Value not exist,*
- *Value is not in interval*, Corresponds to, *Value is invalid,*
- *Value is not equal to*, Corresponds to, *Value is inexact,*
- *Value is not between* (X,Y) , Corresponds to, *Value is unacceptable* and so on.

6. Rules of data anomalies elimination method using for various data values and various data anomalies types – F^C. These rules unambiguously correspond to the relation R^{VUS} between data items, data anomalies, and methods of anomalies elimination.

Ontology constructed in this way may be used as primary ontology for knowledge base of Big Data source cleanup Framework.

When the base ontology is constructed algorithm of Big Data source cleanup can be applied for the first time. The first application may not give an effective result of data cleanup in the general case. It is because the knowledge base need the addition of new knowledge. General steps sequence of this algorithm is like the next.

1. For each data item with anomaly from class C^V , the type of anomaly has to be qualified. Qualification of anomaly means determining anomaly nature and the way of its interpretation. As a result, any concept from category "kind of anomaly"– C^U would correspond to any data item of C^V. If this operation can't be executed complement of ontology is necessary (see step 4 of algorithm).

2. Using data item determination and anomaly type qualification, corresponding relation, and the rule of the anomaly of certain type elimination for certain data item would be defined. Generally, data anomaly elimination rule is the expression of type

$$(V_i \wedge U_k) \rightarrow S_l, \tag{4}$$

where, V_i is data item, U_k is a kind of data anomaly, S_l is method of data anomaly elimination (new data value definition).

3. At the nest step replacement if invalid data value has to be executed according to defined above relations and rules. When whole data resource is processed algorithm to be complete if not – return to step 1.
4. This step is need to recognize and fix the problem situation, appeared during the attempt of data resource cleanup. These problem situations can be categorized according to its origin:

- no description of data item or data value in the ontology;
- no description of anomaly type in the ontology;
- no description of method to eliminate some type of data anomaly;
- no description of rule to eliminate anomaly for particular data value;
- using of rule to eliminate uncertainty did not effect.

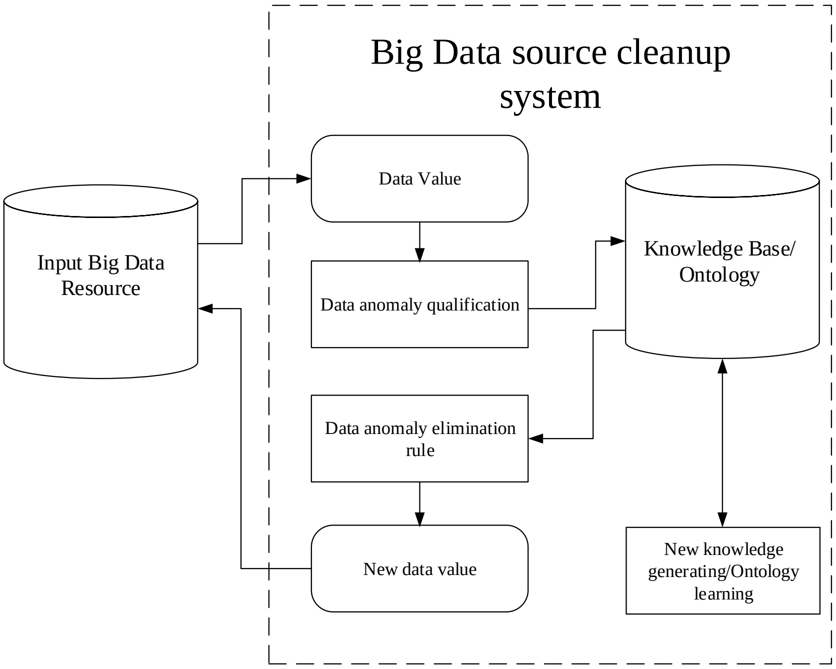

Big Data source cleanup system
Input Big Data Resource
Data Value
Data anomaly qualification
Knowledge Base/ Ontology
Data anomaly elimination rule
New data value
New knowledge generating/Ontology learning

Fig. 3. Data anomalies processing scheme for Big Data resource cleanup

5. If the situation is recognized and categorized, we can fix it by the specified way. These ways are of

- define the new concepts for the data value, for data anomaly type or for data anomaly elimination method;
- define the new item in relation set for definition of correspondence between data value, data anomaly type, and anomaly elimination method;
- define the new rule of data anomaly elimination.

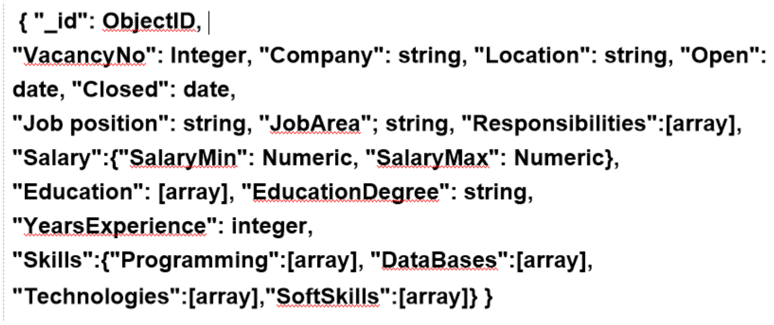
Step 1 of the algorithm needs to be repeated again after the execution of described over operations.

4 The Example

This example has been developed to show how to describe a piece of knowledge about data, data anomalies, and data cleanup methods using ontology tools. RDF-OWL technology [8] together with Protégé ontology editor [7] has been used for this purpose. Big Data resource, which presents the results of the job market monitoring for IT branch in Ukraine, has been considered as the example of developed method of Big Data resource cleanup. Most popular job search web sites such as dou.ua, work.ua, rabota.ua, job.ua have been explored as data sources. The principal values of job search data used as monitoring process dimensions are the next:

- company-employer name,
- company-employer location – city/region,
- work position,
- job area,
- responsibilities,
- salary,
- education specialty/degree,
- job experience,
- necessary skills,
- vacancy duration.

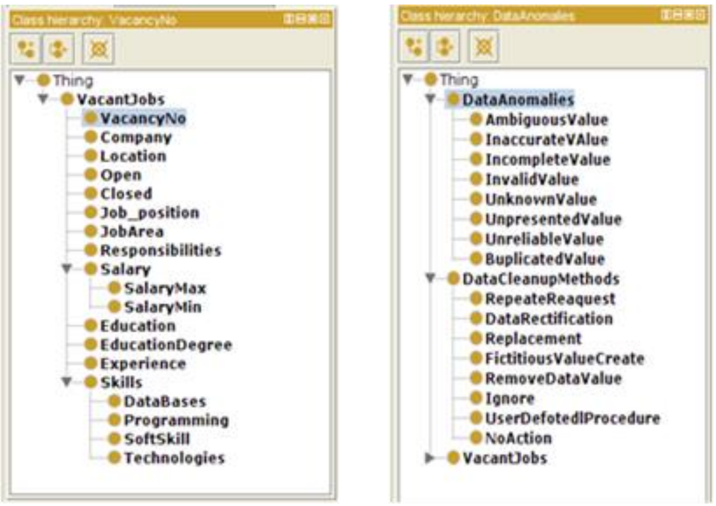
The data set for exploring has been modeled as a NoSQL document-oriented database using MongoDB JSON [11] format (Fig.4).



```
 { "_id": ObjectID,
"VacancyNo": Integer, "Company": string, "Location": string, "Open":
date, "Closed": date,
"Job position": string, "JobArea"; string, "Responsibilities":[array],
"Salary":{"SalaryMin": Numeric, "SalaryMax": Numeric},
"Education": [array], "EducationDegree": string,
"YearsExperience": integer,
"Skills":{"Programming":[array], "DataBases":[array],
"Technologies":[array],"SoftSkills":[array]} }
```

Fig. 4. Data model for Job Vacancy Big Data resource

Primary ontology for cleanup of considered Big Data resource has been developed by Protege ontology editor [7]. Then we have saved the developed one as OWL/XML file. This ontology includes three class of entities: "*VacantJobs*" for data value description, "*DataAnomalies*" for classification of corrupted data, and "*DataCleanupMethods*" for the definition of methods of invalid data processing (Fig.5).

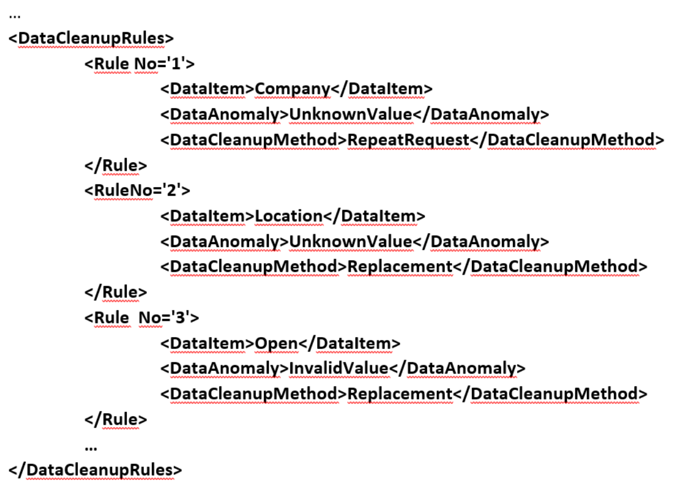


a) Data Values Class

a) Data Anomalies and Data Cleanup Methods Classes

Fig. 5. The structure of primary ontology for Big Data resource cleanup developed by Protégé

The set of rules to describe how certain data values with certain data anomalies has to be fixed by a certain method (during Big data resource cleanup process) has been



```
…
<DataCleanupRules>
          <Rule No='1'>
                    <DataItem>Company</DataItem>
                    <DataAnomaly>UnknownValue</DataAnomaly>
                    <DataCleanupMethod>RepeatRequest</DataCleanupMethod>
          </Rule>
          <RuleNo='2'>
                    <DataItem>Location</DataItem>
                    <DataAnomaly>UnknownValue</DataAnomaly>
                    <DataCleanupMethod>Replacement</DataCleanupMethod>
          </Rule>
          <Rule  No='3'>
                    <DataItem>Open</DataItem>
                    <DataAnomaly>InvalidValue</DataAnomaly>
                    <DataCleanupMethod>Replacement</DataCleanupMethod>
          </Rule>
          …
</DataCleanupRules>
```

Fig. 6. The example of ontology rules for data cleanup in RDF/XML format

developed as well. RDF/XML format has been used for these rules presentation (Fig.6). By this way, we have constructed the primary version of the special ontology described above. Further, it is ready to be improved and developed for efficient support of Big Data resource cleanup processes.

5 Conclusions

An approach to solve the problem of quality of Big Data sources by their cleanup has been considered in the paper. We propose to solve the problem of such type by the development of knowledge-based intelligent tools. The peculiarity of the solution is using an ontology as a core of knowledge base for the Big Data resource cleanup framework. It is the principal difference between the proposed approach and the traditional methods of data cleanup. We consider the ontology as a special type of metadata, which describes Big Data resource, anomalies of data, corresponding methods of data cleanup and the relations between theirs. The developed approach gives us the such possibilities as to design special tools for intelligent Big Data resource cleanup; to make better procedures of data clearing of a Big Data resource; to accumulate for further use of knowledge and experience to solve data quality and data cleanup problem.

The principles and methods developed in the paper may be useful for data scientists at the processes of preparation of Big Data resources to analysis.

Reference

1. Alieksieiev, V., Berko, A.: A method to solve uncertainty problem for big data sources. In: Proceedings of the 2018 IEEE Second International Conference on Data Stream Mining & Processing, DSMP, 32-37(2018)
2. Aliekseyeva, K., Berko, A.: Quality evaluation of information resources in web-projects. Actual Problems of Economics, No 136(10), 226-234 (2012)
3. Date, C. J.: Database in Depth: Relational Theory for Practitioners. O’Reilly, CA (2005)
4. Jaya, M. I., Sidi, F., Ishak, I., Affendey, L. S., Jabar, M. A. : A review of data quality research in achieving high data quality within organization. Journal of Theoretical and Applied Information Technology, Vol.95, No 12, 2647-2657 (2017)
5. Losin, D.: Big data analytics, Elsevier Inc., Waltham, MA, USA (2014)
6. Marz, N., Warren, J.: Big Data: Principles and best practices of scalable realtime data systems, Manning Publications (2015)
7. Protégé. A free, open-source ontology editor and framework for building intelligent systems, https://protege.stanford.edu
8. RDF a Core 1.1 - Third Edition. Syntax and processing rules for embedding RDF through attributes, https://www.w3.org/TR/2015/REC-rdfa-core-20150317 (2015)
9. Rubinson, C.: Nulls, Three-Valued Logic, and Ambiguity in SQL : Critiquing Date’s Critique. In: SIGMOD Record Vol. 36, No. 4, 137-143 (2007)
10. Rusyn, B., Tayanov, V., Lutsyk, O.: Upper-bound estimates for classifiers based on a dissimilarity function, Cybernetics and Systems Analysis, 48(4), 592-600 (2012)
11. Sadalage, P. J., Fowler, M.: NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence, Publisher: Addison-Wesley Professional (2012)