CONCEPTUAL APPROACH FOR PRESENTING TEXT DATA FROM WEB-BASED INFORMATION SYSTEMS IN STRUCTURED FORM

Assoc. Prof. Plamen Hristov Milev, PhD[1]
Yavor Nikolov Tabov[2]

Abstract: Data in web-based information systems is of growing interest for analytical processing by organisations. Usually, the text data in this type of systems has an unstructured form. Unstructured data in web-based information systems has a very large volume. This implies the application of specific approaches to analytical data processing given their natural unstructured form. The need to present text data from web-based information systems in a structured form is highlighted. The article is dedicated to such issues. The peculiarities of the data in web-based information systems are considered. A conceptual approach for presenting text data from this type of systems in a structured form is presented. Based on the proposed conceptual approach, a model of software solution, which is an opportunity for technological implementation of the conceptual approach, is defined. Emphasis is placed on the growing role of data within the web concept of management decision-making in organisations.

Keywords: conceptual approach, text data, web based information system.

JEL: C18, D83, L86.

Introduction

The need to present text data from web-based information systems in a structured form is found mainly in the subsequent possibilities for analysis

[1] Department of Information Technologies and Communications, University of National and World Economy – Sofia.

[2] Department of Information Technologies and Communications, University of National and World Economy – Sofia.

of the text data presented in this way. In the context of raising the level of management decisions made in organisations, it is essential to have data in a form that allows their subsequent analytical processing. In this context, the aim of the article is to perform the following sequence of tasks:

- Clarifying the theoretical statements about the need to present in a structured form the text data in web-based information systems;
- Developing a conceptual approach for presenting text data from web-based information systems in a structured form;
- Defining a model of software solution, which is an opportunity for technological implementation of the conceptual approach.

The implementation of the activities on these tasks creates opportunities to develop software solutions for presenting text data from web-based information systems in a structured form. In this sense, web-based information systems are the object of the study, and the subject of the study is the presentation of text data from this type of systems in a structured form. The aim of the research is to study the features and nature of data in web-based information systems, to develop a conceptual approach for presenting text data from web-based information systems in a structured form and to define a software solution model based on the proposed conceptual approach.

* * *

With the development and extension of the web concept, it turns out that the data within the software solutions that is part of this concept remains inaccessible for analytical processing in its natural form because it lacks the necessary structure to apply existing analytical techniques. However, by identifying key characteristics, data that may seem unstructured may not be completely unstructured (Holmes, 2017). For instance, emails contain structured metadata in the title, as well as the actual unstructured message in the text. Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information (NISO, 2004). On this basis, emails can also be classified as semi-structured data. Metadata labels, which are essentially descriptive, can be used to add structure to unstructured data. Adding a label with words to an image within a web-based information system makes the same image recognisable and easier to search. Semi-structured data can also be found on social networks that use hashtags (descriptive labels), so that messages on a topic that are essentially unstructured data can be identified (Holmes, 2017). The main activities in the process of processing unstructured data within the web concept can be defined in the following sequence:

- Collecting unstructured data;
- Transforming the collected unstructured data into a convenient form for processing;
- Manipulating the converted data;
- Storage of ordered and summarised data and extraction of relevant information from them;
- Output and transmission of the extracted information.

Based on the evolution of the Web concept and the peculiarities of the stages of Internet development, it can be concluded that the variety of opportunities provided by web-based information systems is a prerequisite for the availability of a lot of data with diverse thematic content and lack of centralised control over their form of presentation. If within a web-based information system the data is usually presented in a certain way, then within a group of many web-based information systems the presentation and categorisation of the same data type is usually different for each web-based information system and the whole set of data can be considered as a set of unstructured data. Nowadays, much of the data of interest for analysis by organisations is in an unstructured form within various web-based information systems. In this context, extracting data from web-based information systems is a key activity that is crucial for organisations. According to some research in the subject area, (Das and Kumar, 2013) more than 80% of useful business data is in unstructured form. Using large amounts of data is common for large companies. In Bulgaria, the application of business analysis systems is also intended primarily for large enterprises, while small and medium-sized enterprises experience a number of difficulties in this regard (Shishmanov, 2013). The introduction of digital innovations in companies requires a review of all existing processes in order to make the necessary changes to achieve their digital and corporate goals (Orehov, 2020). Apart from companies in the business sector, data and software solutions for data analysis are extremely important for public sector institutions, where the need to increase the efficiency of e-government is increasingly emphasized (Kirilova and Naydenov, 2021). In this regard, the significance of the computerisation of internal control in the public sector should be noted, in addition to the dependence of this effectiveness on the relevant control and security measures. (Borisov, 2021). Software solutions in general and in particular specialised software solutions for data management are a significant factor in the development of modern economy (Kirilov, 2016). According to some researchers in the subject area, including Khanet et al. (2014), as well as Eberendu (2016), unstructured data is constantly increasing in volume, due to the data that is created daily in modern world-famous platforms with rich user content, such as YouTube, Facebook, Twitter, LinkedIn, etc.

In the context of the possibilities for developing a conceptual approach for presenting text data from web-based information systems in a structured form, it is necessary to make some preliminary assumptions, which have the character of restrictive conditions with regard to the conceptual approach itself:

- The whole process of text data analysis in web-based information systems can be divided into three conditional parts: extraction of text data from web-based information systems; presentation of the already extracted text data in a structured form and analytical processing of the text data presented in a structured form (Fig. 1);
- The conceptual approach for presenting text data from web-based information systems in a structured form aims to establish appropriate methods and technological tools only with regard to the presentation of text data in a structured form. This means that the application of the conceptual approach presupposes the assumption that the text data from the web-based information systems has already been extracted and is in temporary storage;
- The conceptual approach does not imply the inclusion of methods related to analytical processing of text data already presented in a structured form. Analytical processing is the subject of further research outside the scope of this article. In essence, analytical processing may include performing contextual analysis, analysis of specific fragments of the text, etc.

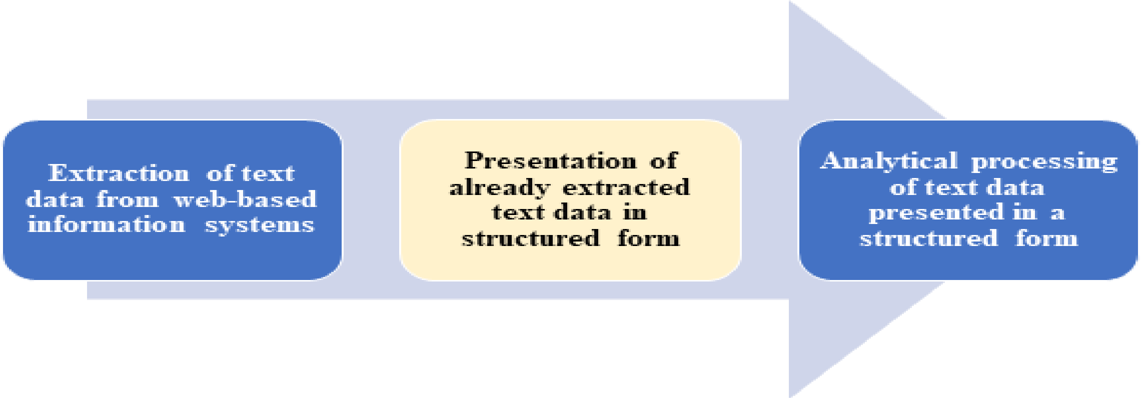
Extraction of text data from web-based information systems
Presentation of already extracted text data in structured form
Analytical processing of text data presented in a structured form

Figure 1. Extraction, presentation in a structured form and analytical processing of text data in web-based information systems

These three assumptions should be considered as restrictive conditions regarding the scope of the conceptual approach being developed. Its main goal is to achieve a correct presentation of the already extracted text data in a structured form, which will end with the recording of the relevant data in a data repository, and this will allow subsequent analysis:

- Method for comparison with a set of templates for presenting text data from web-based information systems in a structured form;
- Method for recognising fragments in text data structures;
- Method for creating a set of templates for presenting text data from web-based information systems in a structured form;
- Method for clearing redundant fragments in text data structures;
- Method for storing configurations for presenting text data from web-based information systems in a structured form;
- Method for visualisation of the text data presented in a structured form.

The proposed system of methods should be precisely specified. In order to make the description of the characteristics of each of the methods clearer, it is good to use a similar method in the specification. The definition of the system of methods imposes and requires both the detailed description and specification of each of the methods in the conceptual approach, and a description of their relationship. For these reasons, and in view of the aim of the article, the following sequence of implementation of the proposed methods is suggested (Figure 2).

The first method in the conceptual approach is a method for comparison with a set of templates for presenting text data from web-based information systems in a structured form. The implementation of this method presupposes the existence of a set of predefined templates for presenting text data from web-based information systems in a structured form.

The second method in the conceptual approach is a method for recognising fragments in text data structures. Before describing the essence of the method, it is necessary to specify that it is the second in the overall sequence of methods in the conceptual approach. Its implementation is carried out in the event that the implementation of the method of comparison with many templates for presenting text data from web-based information systems in a structured form does not give the required result. This practically means that no match of the analysed structure with the predefined structure templates was found.

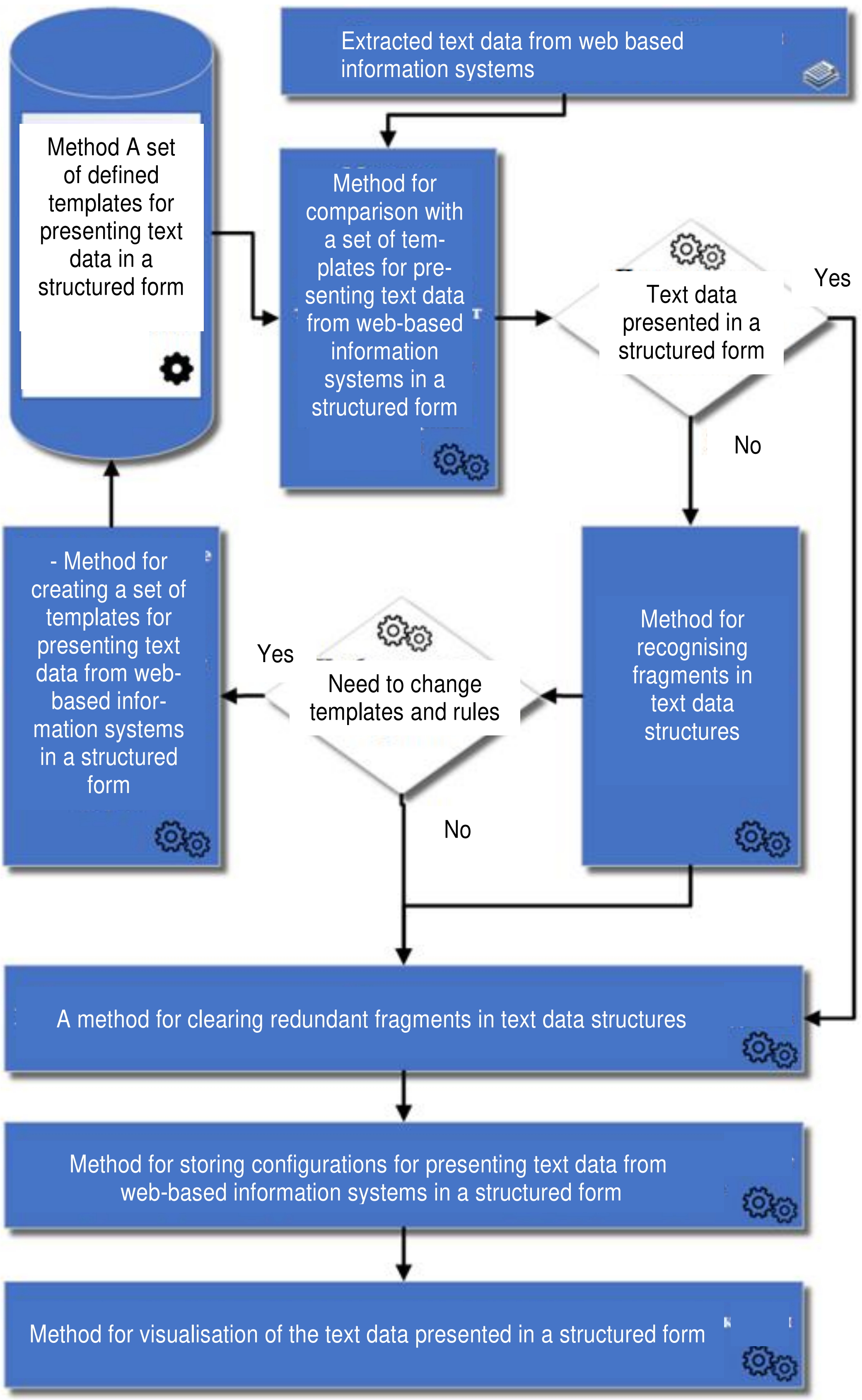
Extracted text data from web based information systems
Method A set of defined templates for presenting text data in a structured form
Method for comparison with a set of templates for presenting text data from web-based information systems in a structured form
Text data presented in a structured form
Yes
No
- Method for creating a set of templates for presenting text data from web-based information systems in a structured form
Yes
Need to change templates and rules
No
Method for recognising fragments in text data structures
A method for clearing redundant fragments in text data structures
Method for storing configurations for presenting text data from web-based information systems in a structured form
Method for visualisation of the text data presented in a structured form

Figure 2. Connection between the implemented methods as part of the conceptual approach

The third method in the conceptual approach is a method for creating a set of templates for presenting text data from web-based information systems in a structured form. The purpose of this method is to provide an opportunity to define different templates for presenting text data in a structured form (in case one of the already existing templates is not recognised and a new definition is required) which should be applied depending on the specifics of the content of a particular web page.

The fourth method in the conceptual approach is a method for clearing redundant fragments in text data structures. The purpose of this method is to provide the ability to clear unnecessary fragments of text data, which will allow the effective presentation of text data in a structured form, depending on the specifics of the content of a particular web page.

The fifth method in the conceptual approach is a method for storing configurations for presenting text data from web-based information systems in a structured form. The purpose of this method is to provide the ability to store configurations for presenting text data from web-based information systems in a structured form.

The sixth method in the conceptual approach is a method for visualising the text data presented in a structured form. The purpose of this method is to provide an opportunity to visualise the presented in a structured form text data from web-based information systems to be exported to external systems.

The description of the proposed methods is presented in Table 1.

Based on the described system of methods in the proposed conceptual approach, a model of software solution for presenting text data from web-based information systems in a structured form can be defined. This model includes the connection between the system of methods and the environment from which the input text data from web-based information systems is provided. The model of this software solution is defined according to the specifics of the classic three-layer architecture of web-based information systems, which consists of a layer of data, a layer of business logic and a layer of user interface (Figure 3).

Table 1
Description of the methods in the conceptual approach

	Aim	Tasks	Result
The first method	To match the web-based structure with a predefined structure template from the corresponding set.	To carry out an analysis of the already extracted text data from web-based information systems; to retrieve a template from the set of defined templates for presenting text data in a structured form; to compare and provide feedback on whether the text structures of web-based information systems correspond to any of the predefined templates.	Found a match with a predefined template.
The second method	To enable automatic identification of fragments in text data structures within a web page.	Defining a new rule for recognising fragments in text data structures within a web page; defining a rule that complements an already existing rule for recognising fragments in text data structures within a web page; defining a rule that modifies an already existing rule for recognising fragments in text data structures within a web page for a specific use case.	A set of recognised fragments in text data structures, which includes a description of the relevant elements and their content in the form of hypertext.
The third method	Creating various descriptions of possible web page structures that can be algorithmically recognised and applied in the form of a template in the respective case of use.	The main tasks of the method are the following: defining a new template for presenting text data from web-based information systems in a structured form; defining a template that complements an already existing template for presenting text data from web-based information systems in a structured form; defining a template that modifies an already existing template for presenting text data from web-based information systems in a structured form for a specific case of use.	A set of templates for presenting text data from web-based information systems in a structured form.
The fourth method	Creating an algorithm for recognising blocks of unnecessary text within a web page that are not of interest for presentation in a structured form	Defining a new procedure for clearing unnecessary fragments within text data structures on a web page; defining a procedure that complements an existing procedure for clearing unnecessary fragments within text data structures on a web page; defining a procedure that modifies an existing procedure for clearing unnecessary fragments within text data structures on a web page for a specific use case.	The creation of an algorithm for recognising blocks of text within a web page that are not of interest for extraction (cleared data).
The fifth method	Creating a repository, where the available configurations for presenting text data from web-based information systems in a structured form is saved.	Defining script to create configuration storage repository for presenting text data from web-based information systems in structured form; creating procedures for recording new configurations for presenting text data from web-based information systems in a structured form; creating procedures for processing requests for access to stored configurations for presenting text data from web-based information systems in a structured form.	Configuration repository for presenting text data from web-based information systems in a structured form.

The sixth method	Visualisation of the presented in a structured form text data, which provides opportunities for export of the text data presented in a structured form to external systems, where relevant analyses can be performed.	Selecting a set of text data presented in a structured form; visualising the selected text data presented in a structured form; exporting the text data presented in a structured form to external analysis systems.	Visualised textual data presented in a structured form that can be exported to external analysis systems.

The implementation of the system of methods begins with the initial configuration, which must include the address of a web-based information system on which the individual methods will be applied. After loading the respective configuration, the text data is extracted from the respective address and the method for comparison with a set of templates for presenting text data from web-based information systems in a structured form is performed. The available templates are loaded and applied to the extracted text data. When finding a matching template, it is proceeded to the method for clearing redundant fragments in text data structures. If no template is found, the method for recognising fragments in text data structures is started. The available rules for recognising fragments in text data structures are loaded. The result of this method is a dynamically generated template or part of a template. After implementing the method for recognising fragments in text data structures, if there are available results, they should be given a configuration that represents the approval or rejection of each of them. Then the method for creating a set of templates for presenting text data from web-based information systems in a structured form is implemented. Within this method, the final template is created in order to be applied in the process of presenting the extracted text data in a structured form. The design of the method for clearing redundant fragments in text data structures includes a visual tool for setting specifics, which in a certain way describes fragments of the recognised structures that must be removed from the final result. Confirmation of the final settings leads to the transition to the method for storing configurations for presenting text data from web-based information systems in a structured form. The implementation of this method is a record of the selected configurations from the previous methods. Then, it is proceeded to the implementation of the method for visualisation of the text data, which shows the practical result of the application of the conceptual approach for presenting text data from web-based information systems in a structured form.

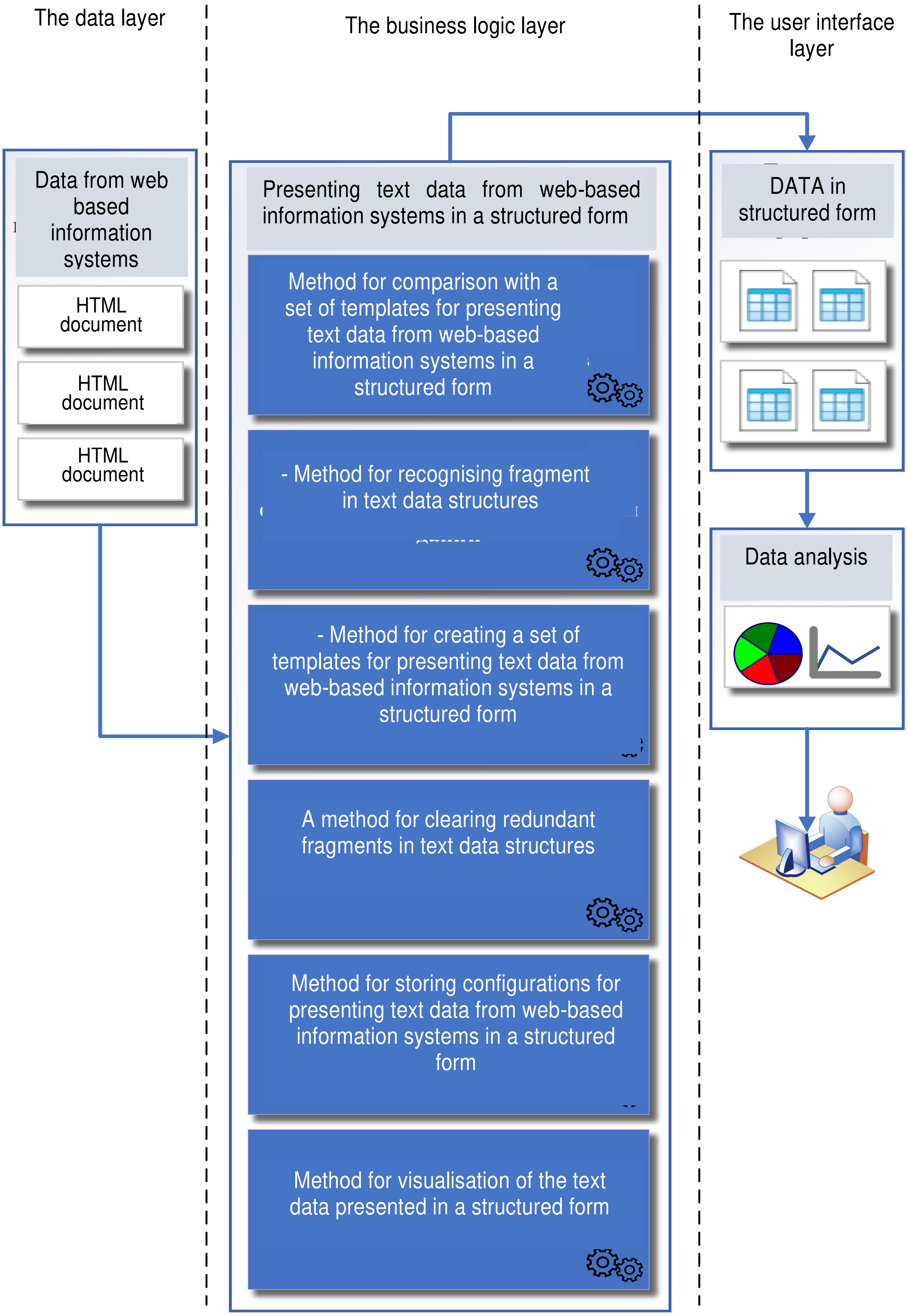
The data layer
The business logic layer
The user interface layer
Data from web based information systems
HTML document
HTML document
HTML document
Presenting text data from web-based information systems in a structured form
Method for comparison with a set of templates for presenting text data from web-based information systems in a structured form
- Method for recognising fragment in text data structures
- Method for creating a set of templates for presenting text data from web-based information systems in a structured form
A method for clearing redundant fragments in text data structures
Method for storing configurations for presenting text data from web-based information systems in a structured form
Method for visualisation of the text data presented in a structured form
DATA in structured form
Data analysis

Figure 3. Software solution model in accordance with the proposed conceptual approach

Based on the research conducted in the article and based on the proposed conceptual approach for presenting text data from web-based information systems in a structured form, the following conclusions can be defined:

- With the development of the web concept, the data in web-based information systems on the Internet is growing significantly, based on which web-based information systems available on the Internet can be defined as the largest number of data systems in unstructured form;
- From a technological point of view, the extraction of data from web-based information systems precedes the application of the appropriate approach for its presentation in a structured form;
- The conceptual approach proposed in the article can serve as a starting point for building software solutions for presenting text data from web-based information systems in a structured form;
- The possibility for analytical processing of the text data presented in a structured form is essential for modern organisations in the context of increasing the adequacy of management decisions.

Conclusion

The main results of the development of this research issue are in the following areas:

- The theoretical statements about the need to present text data from web-based information systems in a structured form are clarified, where the natural form of data does not have the necessary structure to be analytically processed;
- A conceptual approach for presenting text data from web-based information systems in a structured form has been developed. It consists of a system of interconnected methods;
- A model of software solution is defined, representing the practical applicability of the technological implementation of the conceptual approach proposed in the article.

References

Borisov, B. (2021). Opportunities for Modernization and Electronization of Internal Control in the Public Sector. *Business management*, (3), 31-44.

Kirilov, R. (2016). Software Solutions for Managing Projects Co-financed under the European Union’s Operational Programmes. *Business management*, (3), 50-68.

Kirilova, K., & Naydenov, A. (2021). The State of E-government and Digital Administrative Services in the Republic of Bulgaria. *Business management*, (2), 5-21.

Orekhov, M. (2020). The Essence of the Digitalization Process as a New Global Informatization Stage. *Business management*, (1), 75-95.

Shismanov, K. (2013). An Analysis of the Possibilities for the Development of Information Systems in Companies and Organisations. *Business management*, (2), 83-100.

Das, T., & Kumar, P. (2013). BIG Data Analytics: A Framework for Unstructured Data Analysis. *International Journal of Engineering and Technology (IJET)*, *5*(1).

Eberendu, A. (2016). Unstructured Data: an overview of the data of Big Data, *International Journal of Computer Trends and Technology (IJCTT)*, *38*(1).

Holmes, D. (2017). *Big Data: A Very Short Introduction*. Oxford University Press.

Khan, N., Yaqoob, I., Hashem, I., Inayat, Z., Ali, W., Alam, M., Shiraz, M., & Gani, A. (2014). Big Data: Survey, Technologies, Opportunities and Challenges. *The Scientific World Journal*, *2014*.

NISO (2014). *Understanding Metadata*. Bonanza Creek LTER. https://www.lter.uaf.edu/metadata_files/UnderstandingMetadata.pdf

ISSN 0861 - 6604

BUSINESS management

BUSINESS management 1/2022

PUBLISHED BY
D. A. TSENOV ACADEMY
OF ECONOMICS - SVISHTOV

1/2022

Editorial board:

Proofreader – Anka Taneva
English translation – senior lecturer Rumyana Deneva, senior lecturer Radka Vasileva
Technical secretary – Zhivka Tananeeva

The printing of the issue 1-2022 is funded with a grand from the Scientific Research Fund, Contract KP-06-NP3/12 /15.11.2021 by the competition "Bulgarian Scientific Periodicals - 2022".

Submitted for publishinq on 09.05.2022, published on 12.05.2022,
format 70x100/16, total print 40

BUSINESS management

D. A. Tsenov Academy of Economics, Svishtov

Year XXXII * Book 1, 2022

CONTENTS

MANAGEMENT practice