Journal of Electrical Engineering, Electronics, Control and Computer Science – JEEECCS, Volume 4, Issue 13, pages 15-20, 2018

A parallel between a Content Management System for the Internet of Things vs. a Corpus Linguistics Project

Maria Loredana Boca, Mircea Rîşteiu
Department of Exact and Engineering Sciences
“1 Decembrie 1918” University of Alba Iulia
Alba Iulia, Romania
loredana_boca1@yahoo.com,
mircearisteiu@yahoo.com

Liana Boca
Computers Department
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
liana.boca@gmail.com

***Abstract* – This paper, which presents a parallel between two types of Content Management Systems designed for two different domains, try to capture the advantages, disadvantages and importance of the features implemented for this type of application. The main purpose of this paper is to make a parallel between a Content Management System used in the Internet of Things domain (Remote Control of a block heat and power plant, for example) and a CMS created specifically for a Corpus Linguistics Project.**

Keywords - Content Management System, Internet of Things, Computational linguistics, Database

I. Introduction

The analysis and processing of a large number of data has represented and still represents a significant challenge.

In any field, the issue of obtaining, collecting and then analyzing data arises.

The Internet of Things (IoT) is a network of networks consisting of an impressive number of objects / sensors / devices presented in the network, connected through the information and communications infrastructure (and this number continues to grow). [1]

With this in mind, the number of data that results from interconnecting devices or communicating certain information (different sensors) is also very high.

In computational linguistics, domain massive corpora are collected and studied. In body analysis, linguists track and analyze the frequency of certain terms / words, perform registry analysis or analyze collocations. [2]

Extracting valuable information from big corpora requires databases, computational applications specifically designed for linguists, tailored to their needs so that data manipulation is easily done in a short time.

II. The Internet of Things and Big data

A. *The Internet of Things (IoT)*

In recent years, IoT has become a topic that is being treated with the utmost importance in various circles such as engineering, technology, industry, health, automotive, etc.

As mentioned above, the Internet of Things (IoT) is a network of networks consisting of an impressive number of objects / sensors / devices presented in the network, connected through the information and communications infrastructure (and this number continues to grow). [1]

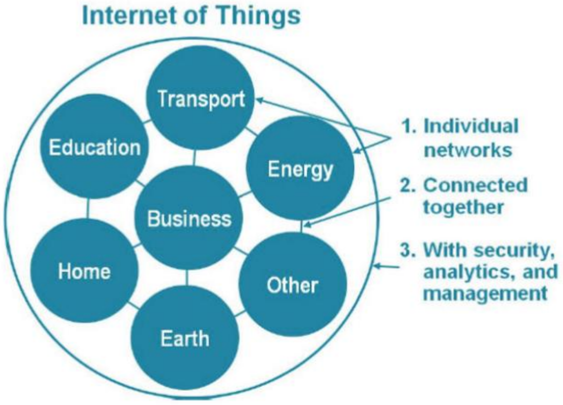

Figure 1: IoT – network of networks [3]

An example of an IoT system may have the following structure:

- data collection - that could be composed of different devices such as sensors, antennas, microcontrollers;
- collate and transfer data - consisting of IoT hub or IoT gateway and IoT Cloud Platform;
- analyze data (take actions if necessary) – user interface, different applications (eg. ERP), Back-end systems. [4]

Note: article presented at the SATEE conference (Smart Applications and Technologies for Electronic Engineering) 21-23th June 2018 ALBA IULIA, Romania

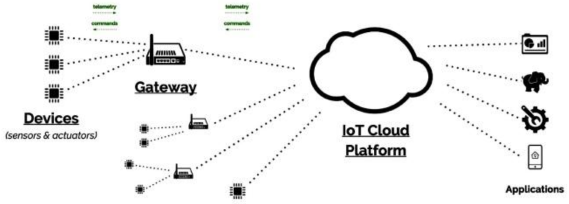

Figure 2: An example of a typical IoT solution [5]

Technological development - the creation, development and improvement of new sensors, mobile and smart devices - which have become smaller and more efficient, with a reasonable purchase cost, improved communications and networks, data storage capacity, “are fueling" the growth of IoT. [6]

Along with the increase in the number of devices, it is logical to increase the number of applications developed for these devices.

IoT is constantly developing, both in terms of increasing the number of devices present in the network and in applications. With this development, perhaps the biggest problem that arises (not being unique) is data integrity: protection, security, privacy and safety.

1) Industrial Internet of Things (IIoT)

The Industrial Internet of Things (IIoT) is a term often used to define various sets of hardware pieces (sensors, devices and machines) that work together through internet of things connectivity to help enhance manufacturing and industrial processes. [7]

The IIoT is not just about replacing old machines / devices that have lagged behind technology with new ones or about the fact that smart machines are more efficient than people in obtaining and then transmitting accurately and in real time the data they get, and neither about negative stories of job losses.

IIoT means the ability to connect or integrate the machines / devices into a system or systems through which the data provided can be monitored, the products can be tracked and the data can be analyzed in real time; most IIoT projects are about process automation, optimization and tactical or strategic goals. [8]

A management system is easier to create, a chain that can be tracked more easily, and in case of need it can intervene in a shorter time. Also knowing and analyzing the data in the shortest time lead to making the best decisions.

B. Data and Big Data

All of these existing devices in the network generate data. Considering the very large number of devices present in the network, it is self-evident that the level of data generated by them is huge.

All of these data should and can be analyzed to get the different information needed in different areas. Of course, data can be saved, and different reports can be made.

“Big Data” is a term that represents the amount of data that is generated by all devices, sensors, objects in the network.

Another definition, known in the online environment and presented in specialized papers for the term Big Data, is that given by Gartner: “high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation”. [9]

This definition is known like the 3V’s: Volume, Velocity and Variety:

- Volume: the amounts of data (extremely large);
- Velocity: speed of change (information is being generated at a rate that exceeds those of traditional systems); [10]
- Variety: the data comes from multiple sources.

Lately, two V’s was added to the 3V’s. The last V’s comes from Veracity and Value:

- Veracity: to cover the trust and uncertainty of information (the data should be accurate, true and reliable);
- Value: the potential for big data to provide a cost-beneficial addition to an enterprise’s technology portfolio [10].

The initial definition known as the 3V’s, subsequently modified to the 5V’s, is now known as the Seven "Vs " of Big Data, and the equations also include the terms of vulnerability and virtue, [21] where:

- Vulnerability starts from the fact that hackers are constantly trying to find safety flaws in large databases to get personal data then used for more or less legitimate purposes; [21]
- Virtue is about the integrity of data and discusses how to collect and then use them. Adopting a proactive transparency policy not only increases confidence [21] but gives organizations the opportunity to present in a positive light the value of the services they offer.

As mentioned earlier in this paper, the level of data generated by the objects / sensors / devices presented in the network is huge, because Big Data are produced in different areas, from intelligent buildings (for safety, equipment, energy-saving), public safety (for traffic lighting, signals, personal identification, emergency), industrial automation – for example block heat and power plant (Remote Control, control options, real time remote monitoring, emergency alarm, water/energy detection), agriculture (environment, greenhouses, soil, irrigation), healthcare (personal identification, equipment, emergency alarm, environment), etc.

C. Corpus Linguistics - Big Data

Corpus is a collection of texts, documents, inscriptions, [11] texts that are part of the "real world" and not only. In [12] the corpus is defined very similarly as a large collection of examples of a particular language, naturally occurring, electronically stored.

In computational linguistics, these stacked text sets are electronically stored and processed. With these corpuses, statistical analyzes, tests, validations of the language rules of a language are carried out.

The first computer corpus, Brown Corpus was created in the early 1960s by W. Nelson Francis and Henry Kucera, at a time when generative grammar dominated linguistics and there was little tolerance for linguistic studies that did not fit into this area. This corpus was therefore not well received by the language community at that time because it is considered that only native speakers of a language can structure legitimate grammatical information about a certain language. However, with the passage of time, linguists have embraced the idea of using corpuscles in their research on languages, both in descriptive and theoretical studies. [13]

The Brown corpus contains over 1,000,000 words extracted from a wide variety of sources. This corpus contains 500 samples distributed across 15 domains, and each sample contains 2,000 words. From the 15 domains we mention: press - reportage (various subjects: sports politics, culture, etc.), press - editorial, religion, general fiction, fiction - science (stories, short stories).

Obviously, such a corpus is information, and it is a very long time to go through it in order to analyze it in the classic version (reading - annotation - analysis - results) by the human operator.

The Brown Corpus was the first computer corpus; today there are many corpuses created, of different sizes, in different languages, corpuses consisting of texts from one or more domains created for different studies.

"Big data" is a term used also in the field of computational linguistics, with reference to the corpus size, of the information that is to be analyzed, which is often huge, the annotated texts and the data analyzed.

III. Content Management System (CMS)

A Content Management System is an interface that allows users to publish content (usually on Web) which provides a simple, accessible website interface that can be used to add content to a page in a highly structured manner. [14] A content management system is not just about publishing content but also about creating, managing, and after publishing - archiving it.

Functions and features of a CMS can vary from a CMS to another, but in general, the main are:

- indexing data
- search and retrieval information
- format management
- revision control
- publishing the information that is wanted. [15]

Within the functions mentioned above, we also consider that the following functions are very important for a CMS:

- storing data
- access control
- reporting.

In recent years, large amounts of money have been invested at companies to buy CMSs that ultimately have failed to provide the necessary functionality. This has led to the creation of new CMSs or the modification and / or adaptation of existing ones to meet the requirements and needs of end users.

Using CMSs specifically tailored to the needs of companies has increased in recent years due to their usefulness, streamlining management operations, increasing the accuracy of data usage, functionality, and last but not least because they have helped to reduce human operator errors.

IV. CMS for a Corpus Linguistics Project

A. The Corpus Linguistics Project

This article is part of the project *Universals and variants of English and Romanian business metaphors. A corpus-based conceptual mapping of contemporary journalese from a pedagogical approach* (University of Alba Iulia, Romania)

Within this project, the team has created two corpora consist of articles that appeared in the following newspapers: The Economist, The Guardian, The New York Times and The Telegraph for the corpora in English and Adevărul, Jurnalul Național, Cotidianul, Capital and Ziarul Financiar for the Romanian corpora, each corpora sum totaling over 500,000 words for each of the two languages.

B. CMS for UvaBu Met Project

Starting from the sizes of the corpora that needs to be analyzed not only by one person but by the team located in different places in Europe, there has been a need to create a data storage and access system for adding, editing and if it is necessary to delete them, this system has been designed, creating a database to be accessed by team members through a Content Management System callable from a web browser.

The need to use a CMS occurred when:

- there was a great deal of information of the same type to be structured and analyzed;
- team members were to have access to the same data stored in a single database;
- team members were required to structure data according to the same criteria;
- information management was to be carried out by a group of people with knowledge of the use of IT applications and not their development.

This Content Management System was designed to provide all the necessary features for all team members to manage the created corpuses without the need for HTML or PHP knowledge.

The CMS has been designed to be easy to use, and the linguist's requirements for information to be added and then managed have been achieved.

The database is stored on a server. With CMS, authorized people logged on with credentials have access to the database and can manage the information that's already stored, or enter new information.

Given that the work of linguists began by creating the two corpuses and then continued and continued by going through the texts (newspaper articles), the annotation of the metaphors and their fitting into one of the three types of metaphors: conceptual, lexical and cultural metaphors, the CMS has been structured so that all the data for this process can be entered in the database, then there are options for managing these data.

As we said, it all starts from adding a new article (part of the corpus) to the database.

The following fields can be completed for any article:

- title
- subtitle
- release date
- content.

Each article has one or more authors (for each author the name, surname and various information fields can be filled in) and each article appeared in a particular newspaper for which names and information fields can be filled in.

The metaphor's text is annotated; each metaphor is part of one of the three types (conceptual, lexical or cultural metaphors) and can be classified into a particular category and / or subcategory.

A search area is available in CMS for the members of the team, with which different information can be found and displayed: searches for a particular metaphor, a lemma, a category or subcategory, an article or an author can be made.

V. CMS for the Internet of Things

The architecture and design of a CMS used in the IIoT was thought and then implemented specifically for each industry. Thus, CMSs have been created specifically for the oil and gas industry, the energy industry, the automotive industry or the manufacturing industry.

An IoT device could send data to a CMS and that way we could have Remote Control of a block heat and power plant, for example, or for a photovoltaic power station. [19]

In a CMS created for a power plant, the sensors initially sent information on voltage or frequency, the parameters that were recorded for later analysis. With the passing of time, the development of various devices as well as applications, the information sent by wireless sensors in the application, are used in real time to identify different events that occur within the network to maintain its stability. For example, the response of the system to a lightning strike is quite different than a transformer failure, [16] and starting from presenting the data in real time and then interpreting it, the system can automatically perform certain operations for its stability (in case of need it will display an alert for the human operator) or provide the operator with the necessary options for making the final decision.

Different data, such as real time remote monitoring of the power generator parameters or control options such as switch On/Off the different parts of the device, alarm and notification system are now parts from a block heat and power plant CMS.

European Standard EN 13306 (2001) defines maintenance as the combination of all technical, administrative and managerial actions during the life cycle of an item intended to retain it in, or restore it to, a state in which it can perform the required function. [17]

The role and importance of maintaining the system under all circumstances is obvious. A CMS created specifically for the IIoT must provide the user with various options for maintaining the system properly, but also for its preventive maintenance. Thus, maintaining the optimal parameters of the whole system is one of the functionalities that a CMS has to offer.

VI. The Parallel Between a CMS for the Internet of Things vs. a Corpus Linguistics Project

After all the above-mentioned information, the common and different functionalities that a CMS created specifically for the Corpus Linguistic Project or the IIoT are obvious. However, for a better image of these functionalities, we tried to structure them in the table below.

Domain Functionalities	Corpus Linguistics Project	Internet of Things
system login based on credentials	✓	✓
data sharing	✓	✓
input of data	of human staff	data are retrieved by sensors
accuracy of data accessed	very high	very high
storage of data on server	✓	✓
real time remote monitoring	unnecessary	✓
rapid access to information	✓	✓
alarms in case of need	unnecessary	✓
dedicated search area	✓	✓
notification system	✓	✓
maintaining the system	✓	✓

reports	✓	✓

Table 1: Functionalities that a CMS created specifically for the Corpus Linguistic Project offers them in relation to those offered by a dedicated CMS for IIoT

The login systems based on credentials as well as information sharing and use by more people (who are mostly trained to use the system) are common points for both CMSs created specifically for a Corpus Linguistics Project as well as for Internet of Things, and vital for data integrity and security.

For Corpus Linguistics Project, data entry is performed by the human operator. In the case of a CMS made for IoT, input data are retrieved by sensors.

Data entered in a CMS for a Linguistics Project is saved in the database that is stored on a server. In the case of IoT, the data transmitted by the sensors / obtained by them, before being saved, can be subjected to filtering processes to enhance their reliability and to convert raw data into a more convenient format for further processing, [17] to detect the malfunctioning data and the intrusion data. [18]

As we said earlier, storing data entered in a CMS for a Language Project is easy, in a database stored on a server. In terms of data storage in IoT it is much more complex, requiring some additional steps, the network architecture in this case being much more developed, containing gateways, routers, cloud and / or a data filtering system. [18]

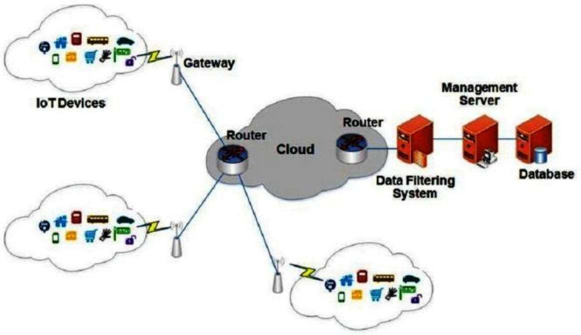

Figure 3: An example of IoT system [18]

Due to the type of system, the way of structuring and manipulating the data as well as the level of data processed, the periodicity of the request must ensure that the system maintenance can be higher (a few weeks) in the case of a CMS for a Linguistic Project, (a few days - maximum) for a CMS for IoT.

Acknowledgment

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS – UEFISCDI, project number PN-II-RU-TE-2014-4-2785.

Conclusion

Within the Corpus Linguistics Project (UvaBu Met Project), a database was created and an entire system for studying texts from business language and metaphor analysis was thought and then implemented (according to the linguist's requirements). A Content Management System was designed to provide all the necessary features for all team members to manage the created corpuses without the need for HTML or PHP knowledge. This CMS has a simple, clear, easy-to-use interface so that the content (articles in the business papers that make up the two corpuses) is easy to manage.

Different CMSs and other applications have begun to be developed and increasingly used in recent years in both IoT and in different industries (IIoT). Companies have become more and more interested in IoT and what they can offer has led to the need to create standards and frameworks. The parallel between CMSs implemented for the two so different areas has been achieved with a number of very important security, data entry and storage points, notification and reporting.

Although the domains in which such applications are used are so different, there are some common points, as shown in the table above.

Future studies will be focused both in the area of computational linguistics for the development of the application needed by linguists for both metaphor and IoT to analyze existing applications developed for different areas of the industry in order to improve them and provide users / operators with practical solutions, easy to use and reliable.

References

[1] **Perera, Charith, et al., et al.** Privacy of Big Data in the Internet of Thing Era. *IT-Pro.* May/June 2015. https://pdfs.semanticscholar.org/6e42/fd9267e2a10aa7783645094cb45d894b377a.pdf.

[2] **Grazib, Mohamed**, *ELECTRONIC CORPORA: AS POWERFUL TOOLS IN COMPUTATIONAL LINGUISTIC ANALYSES.*. Saida : CEUR Workshop Proceedings, 2009. Conférence Internationale sur l'Informatique et ses Applications 2009. Vol. 547. Algeria. ISSN.

[3] **Evans, Dave.** The Internet of Things - How the Next Evolution of the Internet Is Changing Everything . *https://www.cisco.com.* [Online] 2011. https://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf.

[4] **Rouse, Margaret, et al., et al.** Internet of Things (IoT). *https://techtarget.com.* [Online] https://internetofthingsagenda.techtarget.com/definition/Internet-of-Things-IoT.

[5] **Eclipse IoT Working Group.** The Three Software Stacks Required for IoT Architectures. *https://iot.eclipse.org/.* [Online] 2016, 2017. Available under the Eclipse Public License 2.0 (EPL-2.0). https://iot.eclipse.org/resources/white-papers/Eclipse%20IoT%20White%20Paper%20-%20The%20Three%20Software%20Stacks%20Required%20for%20IoT%20Architectures.pdf.

[6] **IEEE Standards Association.** IoT_Ecosystem_Study_2015. *www.sensei-iot.org.* [Online] 2015. http://www.sensei-iot.org/PDF/IoT_Ecosystem_Study_2015.pdf. ISBN 978-0-7381-9501-8.

[7] **Janalta Interactive Inc.** Industrial Internet of Things (IIoT). *https://www.techopedia.com.* [Online] https://www.techopedia.com/definition/33015/industrial-internet-of-things-iiot.

[8] **i-scoop.** The Industrial Internet of Things (IIoT): the business guide to Industrial IoT. *www.i-scoop.eu.* [Online] i-SCOOP. https://www.i-scoop.eu/internet-of-things-guide/industrial-internet-things-iiot-saving-costs-innovation/.

[9] **Gartner, Inc.** IT Glossary - Big Data. [Online] Gartner, Inc. https://www.gartner.com/it-glossary/big-data.

[10] **O ' Leary, Daniel E.** *' Big Data', The ' Internet of Things' and The 'Internet of of Signs'.* s.l. : John Wiley & Sons, Ltd., 2013, Intelligent Systems in Accounting, Ffinance And Management, Vol. 20, pp. 53-65. Published online in Wiley Online Library (wileyonlinelibrary.com).

[11] **Academia Română, Institutul de Lingvistică "Iorgu Iordan".** Corpus. *Dicționarul explicativ al limbii române.* 2nd. s.l. : Univers Enciclopedic, 1998.

[12] **Bennett, Gena.** Using Corpora in the Language Learning Classroom, Corpus Linguistics for Teachers. *www.press.umich.edu.* [Online] 2010. University of Michigan. http://www.press.umich.edu/titleDetailDesc.do?id=3715.

[13] **Meyer, Charles.** *English Corpus Linguistics - An Introduction.* Cambridge : Cambridge University Press, 2002. ISBN 0 521 80879.

[14] techopedia - Where IT and Business Meet. [Online] https://www.techopedia.com/definition/24075/content-management-system-cms.

[15] **Rouse, Margaret.** content management system (CMS) . *TechTarget.* [Online] https://searchcontentmanagement.techtarget.com/definition/content-management-system-CMS.

[16] **Mann, Jason.** Internet of things applications across multiple industries. *www.sas.com.* [Online] https://www.sas.com/en_us/insights/articles/big-data/internet-of-things-applications-across-industries.html.

[17] **Juselius, Juha.** *Advances Conditon Monitoring Methods in Thermal Power Plants.* LUT School of Energy Systems, Energy Technology, LAPPEENRANTA UNIVERSITY OF TECHNOLOGY. 2018. Master's Thesis.

[18] **Kim, Dae-Young, Jeon, Young-Sik and Kim, Seokhoon.** Data-Filtering System to Avoid Total Data Distortion in IoT Networking. *symmetry.* 9, 16, 2017.

[19] **Uhlig, Jens.** Berlin Innovations. *Innovative Produkte, Verfahren und Dienstleistungen.* [Online] https://www.berlin-innovation.de/uploads/tx_innodb/182-m2mgo-showcases.pdf.

[20] **Rouse, Margaret, et al., et al.** Industrial Internet of Things (IIoT) . *https://techtarget.com.* [Online] https://internetofthingsagenda.techtarget.com/definition/Industrial-Internet-of-Things-IIoT.

[21] **DeAngelis, Stephen.** The Seven "Vs" of Big Data *www.enterrasolutions.com* [Online] https://www.enterrasolutions.com/blog/the-seven-vs-of-big-data/