# A Text Analytics Architecture for Smart Companies

Mariagrazia Fugini(*), Jacopo Finocchi(*), Filippo Leccardi(**), Paolo Locatelli(**), Alfredo Lupi(^)

(*) DEIB - Politecnico di Milano, Milano, Italy
(**) DIG - Politecnico di Milano, Milano, Italy
(^) Microdata Group, Cremona, Italy

mariagrazia.fugini@polimi.it, jacopo.finocchi@i3lab.net, filippo.leccardi@polimi.it,
paolo.locatelli@polimi.it, a.lupi@microdatagroup.it

***Abstract* — This paper presents an architecture for Big Data Analytics regarding unstructured content. The architecture is proposed as an industrial solution in a real setting. In particular, the paper focuses on *BD (Big Data) for Smart Companies* and on E*nterprise Content Management* for extraction of information from textual BD. It presents the process architecture and discusses some experiments done as part of a larger BD Analytics project.**

***Keywords — Context Aware Computing, Text Analytics, Big Data, Enterprise Content Management, Natural Language Processing.***

## I. INTRODUCTION

This paper is an outcome of the SIBDA (Sistema Innovativo Big Data Analytics) Project[1] concerning the study and development of an integrated solution in the field of *BD Analytics (BDA),* for three companies, which created an “agreement for competitiveness”. The results show the potential of BD tools and methods to the realization of a *Smart Company*.

We are witnessing a growing relevance of *unstructured information for enterprise business processes*, compared to structured data, as in traditional database information. Smart companies will have to develop effective solutions to enable knowledge extraction from and about text documents. Among the main objectives of works and research in this area, there is the creation of Smart Platform models for data collection and management, which should be *replicable*, leading to standard solutions suitable for enterprise contexts.

In the SIBDA project, we tackled *Enterprise Content Management* (ECM) [1]. This consists in the evolution from classic document management applications towards advanced *content management applications*, characterized by an important role of semantics also in the enterprise environment, and by the diffusion of BD storage tools, such as document-oriented databases. In this scope, a relevant trend in the analysis of unstructured data is the application of Natural Language Processing (NLP) techniques - frequently based on Machine Learning (ML) - to draw from textual data their semantic value and use it to enhance search functionalities or to develop new value-added services.

The SIBDA Project aims at integration of the activities and assets of the three involved companies:
a) *Mailup*, in charge of the *Data Warehouse* and of the general purpose BDA tools;
b) *Microdata Group*, specialized in outsourcing services for management of *document processes*, developed the automatic document analysis tools;
c) *Linea Com* (now *A2A Smart City*), in charge of creating an IoT platform for the acquisition of sensor data.

The paper focuses on the ECM issues under a context-aware approach, to extract knowledge from documents with the flexibility needed to handle the contents of different customers.

The *novelty of the approach* is twofold: treating documents as unstructured BD and immerging ECM in a context-aware environment, where knowledge about the context where textual data is mined out, characterizing the background where BD are generated and used. Examples of information regarding context are: who they are about (customers), in which process they are used, in which domain, and for what procedure. The purpose of *environment knowledge* extraction is to understand the factors that influence the way documents are handled, distributed, archived, etc., with the aims of: i) elaborating better strategies of document management and archiving, and ii) of offering new services to the customers.

The paper presents the system architecture and discusses some experiments done as part of the project.

## II. RELATED WORK

Tools to manage large amounts of heterogeneous data *in real time* are increasingly important to address the new sources of data, such as devices and sensors of the Internet

[1] “Big Data Project” funded by Regione Lombardia with LineaCom, MailUp, Microdata and Politecnico di Milano. The project, with a duration of 2 years, was completed December 2017.

of Things (IoT), social media web platforms, digital media and multimedia systems and business processes. The most significant techniques and technologies are discussed in [2]. Many works and ideas are currently available in the literature.

In [3], context aware software infrastructures are presented as a way not only to receive context information, but also to develop services that may be customized according to user needs. Analogously, our solution is based on an enterprise service bus to process contextual information and offering new services to the customers of one partner (Microdata Group).

[4] is a survey about "intelligence" in IoT systems through context awareness. Context-aware computing is illustrated as requiring both sensing and learning capabilities as IoT systems get more data and better learning skills from the BD sets. This survey reviews the field, covering ubiquitous and pervasive computing, ambient intelligence, and wireless sensor networks, and then, move to context-aware computing studies. Finally, learning and BD studies related to IoT are reviewed. [5] considers how data sensing, information processing, and networking technologies are being embedded in cities to enable the use of innovative solutions towards sustainability and urbanization. Driving such transition predominantly are BDA and context-aware computing and their increasing amalgamation within a number of domains. The work develops and evaluates the most relevant frameworks pertaining to BDA and context-aware computing in the context of smart sustainable cities. It brings together research directed at conceptual, analytical, and overarching levels to stimulate new ways of investigating their role in advancing urban sustainability. We argue that BDA and context-aware computing are prerequisite technologies for the functioning of smart sustainable cities.

In [6], text analytics (or text mining) is presented as one of the major fields of BD because most of the data are stored is in the form of text. Textual data can be unstructured due to different formats (and contain many relevant pieces of information, which, if effectively analyzed, can result in efficient decisions. Analyzing vast amount of textual data will lead to effective results in every business domain and will enable many industries to utilize this paradigm for interpreting their customers' needs. We also consider the growing interest in textual data mining from different business domains. This interest calls for a coherent and intuitive method in BDA to discover hidden patterns and correlations, and to monetize the data to improve the services, particularly for ECM.

[7] combines the interest for context-aware computing systems and the area of treatment of huge amounts of data. An overview of context-aware computing is given; BD challenges are outlined for each stage in the cycle where data are acquired and processed to derive context. A map of existing context-aware systems that handled BD in different ways. Documents as BD are not mentioned, since this is a very new topic in BD management, as we discuss in this paper.

In [8], we described the needs and solutions of the three involved companies, which led to define an overall framework in the field of BD through the description of application cases. The functional and technological requirements of the modules of an integrated BD architecture are given.

## III. TRENDS IN UNSTRUCTURED INFORMATION AND BIG DATA

We are seeing a growing relevance of unstructured information for the enterprise business processes, compared to structured data, like the traditional database information [9].

The spread of BD has changed the type of data collected and processed by organizations' information systems. In recent years, the "BD Analytics and Business Intelligence Observatory" of Politecnico di Milano has recorded a progressive increase in unstructured data and data coming from outside the company's information systems in Italy. The Observatory has studied this tendency year by year and, as reported in the 2017 Research Report "*BD is now: tomorrow is too late*" [10], the increase is continuing in 2017 with the percentage of external data collected by company systems rising to around 21%, compared to 16% three years ago. Congruently, by referring to 3Vs introduced by Doug Laney in his definition of Big Data, the variety of data is identified as the most valuable feature of BD by 43% of organizations, while 32% prize the volume, and for one in four companies the most outstanding feature is the speed of analyses. However, against this backdrop of growing heterogeneity of collected information, the capacity of organizations to grasp the opportunities offered by the data is not growing at the same rate: if we look at the percentage of data effectively analyzed compared to the past we see a drop from 52% to approximately 40% of all data collected, demonstrating a widening gap.

The gap of opportunities is also shown by the interest that each year is reported among the Italian CIOs and Innovation Managers, thanks to a survey among those who belong to medium-large Italian businesses, in Business Intelligence, BD and Analytics. This survey reports no signs of waning [11]. For 2018, the fourth year in a row, it is still the leading priority for investment, capturing the interest of 43% of the sample, ahead of digitalization and dematerialization of processes and documents (35%), consolidation of applications, development and updating of management systems and ERP (29%), and Information Security, Compliance and Risk Management systems (28%). Over the years, the "BD Analytics & Business Intelligence" Observatory has documented how companies have gradually moved toward BD Analytics: from initial confusion to the pursuit of a data-driven strategy, to final awareness of the importance of accompanying the technical innovations with an organizational model fit to manage the transformation. Those, who were first to move beyond the uncertainty, today boast a portfolio of more efficient and effective processes, new products and services, with a certain and measurable return on investment. According to the Observatory estimates, the Analytics market continued to grow during 2017, recording 22% growth for a total value of €1.103 billion that, again, witnesses that the opportunity gap is still open and that players that moved first are keeping on investing. This dynamic growth has also another aspect that needs attention. There are two main sources in the experienced growth. Firstly, large companies who are already familiar with the opportunities offered by descriptive data analysis are studying new projects focusing on forecasting aspects and engineering algorithms aimed at automating processes and services, in pursuit of what we could call the second wave of the data-driven strategy. Secondly, small and medium businesses' interest in data analysis is generally increasing with the adoption of tools of data visualization and basic analysis and of services to support marketing activities. The data show that differently from the past, today the use of BD Analytics is imperative to avoid the risk of losing competitive edge and missing opportunities, and, in extreme cases, of being excluded from new markets, or even from current ones. BD Analytics, initially the prerogative of a few far-sighted, proactive companies, has proven to be a successful innovation, and has become a necessity for survival in competitive markets.

## A. *The Big Data Journey Framework*

The "Big Data Analytics & Business Intelligence Observatory" at Politecnico di Milano has developed a framework called *BD Journey* whose goals regard the support to organizations in understanding the main aspects to be considered when the company wants to evolve from being a "Traditional Enterprise" to a "BD Enterprise".

The BD Journey is a framework that witnesses the complexity and the broadness of impacts of BD Analytics projects. The technological aspect is just one of the four dimensions included in the framework that are the following ones:

1. Strategy;
2. Data;
3. Competence & governance;
4. Technology.

Five intermediate steps compose each item. Steps represent the evolution needed from a traditional approach to a mature BD approach, which the companies involved in a BD Project have to face.

The first dimension, namely *Strategy*, refers to the *organizational approach to Analytics* in the medium-long term. The second dimension concerns *Data* and refers to the way the organization stores data and makes them available. The third dimension analyses the *Competences* needed in the organization to manipulate and analyse data, and regards the analytics "Governance" structure. The fourth and last dimension, namely *Technology*, focuses on the infrastructural approach to BD Management.

## B. *The Big Data Journey applied*

One of the goals of BD Journey is to apply it to assess the maturity level reached by single companies and to estimate the average maturity at national level. The BD Journey has been applied to the BDS developed in the SIBDA project, using two different approaches: first oriented to evaluating the maturity model obtained at the end of the project; the second aimed at assessing which maturity level could be reached in 2/3 years. The resulting positioning is described in the following considering each dimension.

### *Strategy*

Considering this dimension, the BDA solution received a positioning evaluated between the Tactical and Planned steps. BD opportunities are known at Top Management level and there is a plurennial plan that involves BD Management. The planned step is not fully achieved because it is not yet possible to talk about shared KPIs. Each of the companies involved in the project has contributed in developing the BDA solution coherently with its own strategy. In 2/3 years, the Planned step can be fully achieved through the definition of KPIs, for periodical evaluation of activities based on data analysis and enabled by the new DWH.

### *Data*

Considering this second dimension, the step reached by the BDA solution is within the Available step, because data are centrally collected but each Business Unit among those, which actually migrated their own data into the DWH, accesses the DWH thought different solutions and tools based on its own needs. In 2/3 years, the expected position might be harmonized if the data of all Business Units will be moved to the DWH and if also unstructured data will be stored therein. This evolution will enable other analysis, which possibly brings new benefits.

### *Competence & Governance*

Regarding this third dimension, the positioning is at the beginning of the Shared step, because the main competences related to BD management are limited in the IT department of the MailUp Company. Whenever other Business Units might need to access to BD related services, from the extraction of data to the analysis they need to rely on the competences of IT. In 2/3 years, the expected position is far from the actual because Competence & Governance dimension is where the biggest improvements can be made. The expected position at the Data Science Driven step will be reached if there will be a periodical coordination plan focused on the identification of needs and if will be structured a competences development process with the introduction of data scientists that will be part of a cross functional team.

*Technology*

The last dimension is positioned at the Predictive step because the DWH cannot be considered traditional but the analysis conducted area mainly traditional, the first steps have been moved toward an advanced use of analytics. The predictive step is still not fully achieved, because not all the exploitation of predictive analysis has been completed. In 2/3 years, the expected positioning is going to be on the Pattern Identification step, which means that all the potentialities of predictive analysis will be exploited, and more advanced techniques will be used in specific areas, e.g., in fraud detection. The further step (Real Time) is not seen as needed, even if technologically feasible.

The position of the BDA system developed within SIBDA is reported pictorially in Fig. 1.

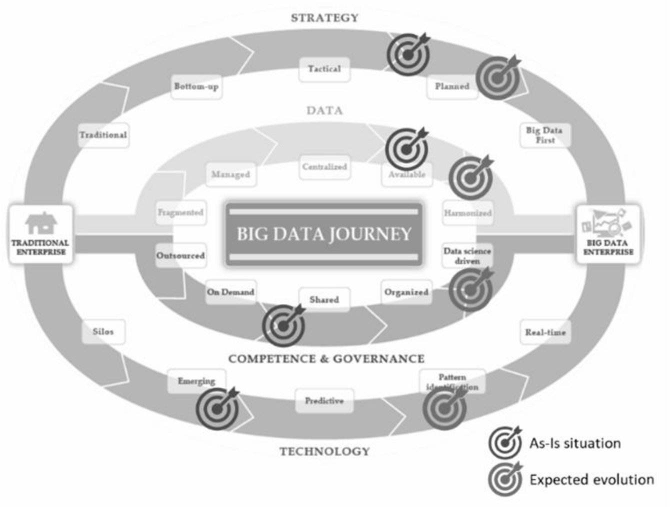

Figure 1 Positioning of the SIBDA BDA system in the Big Data Journey framework

## C. Data as a Service: first signs of innovation

Since the definition of Doug Laney, the word Variety has been closely linked to the concept of BD. The Observatory of Politecnico di Milano has deepened the types of data analyzed by large organizations, in order to assess whether Variety is still an obstacle or not, and the sources of these data, paying particular attention to the ability of large companies to seize opportunities for monetizing information [9].

With regard to the type of data used, the results are not very encouraging: in the Italian scenario, structured data, transactional and non-transactional, continue to account for 84% of the total of data used, without any change compared to 2015, the last year in which this information is was detected by the Observatory. This picture shows clearly that Italian large organizations still do not exploit the potential of unstructured data. Inside this scenario, there are two clusters of organizations: those, which have understood the value within data by buying or selling them (60%) and those, which have not yet explored these possibilities (40%). By referring to the first group of organizations, emerged a decrease in the weight of structured data (79%), mainly due to a fall in the weight of transactional data (-6%). This decrease is in favor of other types of data, primarily geo-spatial data, which weighs 7% of the total (compared to an average of 5%), and unstructured textual data, e-mails or documents, which reach 13% (compared to an average of 10%). The situation is totally reversed for that 40% of large organizations that declare to have no relations with the outside in the exchange of information. These companies show a considerable delay in the use of data coming from various sources, - the weight of structured data reaches 93% - and they remain lagging behind in an international scenario that continues to affirm the disruptive potential of unstructured data.

## D. Towards Machine Learning, Deep Learning and Real Time

The Observatory identified two directions of innovation: on one hand the techniques of Machine Learning and Deep Learning, which enable new types of analysis; on the other the Real-time Analytics.

Regarding the use of Machine Learning and Deep Learning in the development of Analytics projects, it is interesting to note that 62% of large companies claim to have specific skills. Just over a third of these have already been internalized and a further 30% plan to do so within the next two years. Companies that already have these skills within their own show a higher level of maturity in the implementation of projects, both in the number with an average of seven initiatives, both from the point of view of functional coverage, which is wide on many business processes.

Regarding the Real Time analysis, this means setting up a technological infrastructure capable of processing and making data available in a timely manner and using tools for analyzing and displaying information that are able to interface with a dynamic environment. 11% of organizations today exploits methods of analysis in Real-Time, where data is collected in real time and can be interrogated when it is needed, or in streaming, where instead there is a continuous flow of collection data that must be analyzed continuously. An additional 33% of companies have an infrastructure capable of enabling analysis in Near Real-Time mode, therefore with a refresh rate that is reduced under the hour (generally 15 or 30 minutes). Finally, only 56% of organizations can analyze data only in Batch mode, then with a system update at pre-defined and wider time intervals, usually daily. The comparison with last year shows a significant increase in the number of organizations that perform analyzes at least in Near Real-Time mode: from 26% to 43%.

## IV. SYSTEM ARCHITECTURE

The overall BDA system was designed as an integration of many components and assets characterizing each company, including hardware and software, skills and collected data. Now we focus on that portion of the overall system consisting in a highly automated ECM architecture, shown in Fig.2, as an evolution of the pre-existing Document Management System adopted by *Microdata Group*.

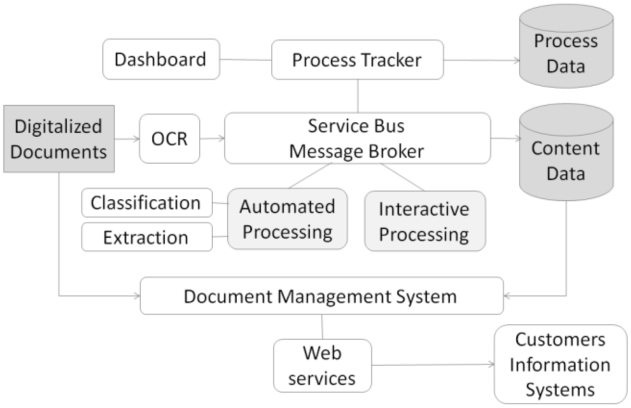

Figure 2. Schematic view of the ECM System.

*Microdata*, operating in the document storage outsourcing market, deals with systems designed to process a flow of incoming documents from multi-channel sources, to identify and extract, from each document, an information set and finally store the data and metadata about data processing, together with the documents.

In particular, *Microdata* aimed to improve the speed and the efficiency of the documental data processing cycle, by increasing the automation level without losing quality and possibly providing a near real-time content analysis. It also wanted a solution to extract content data from the documents, and to make such knowledge available to its customers, possibly providing new services.

The scope of this portion of the system fully falls into the BD field, as it requires processing a large volume of information, at a high speed (in some cases in near-real time) and above all dealing with the high variety and complexity of the data, which consists of non-structured text.

The design of the ECM system takes into account the need to extract information from documents with a structure that is not described formally a priori, carrying out a complete extraction of all the textual information contained in the document.

The implementation choices adhere to the following directions: choosing components that can be integrated into an industrial process, within a server-based solution (avoiding the need to install software components on client stations and minimizing the operators' interventions) and minimizing the specific implementation effort to be dedicated to each documental process or to each document type, exploiting tools with machine-learning capabilities (which progressively improve the quality of the result as the number of cases processed increases).

The automated solution, when operational, will cover most of the company document processing activities, although some non-automated processes will be left, where human operators interactively extract and store the document content. An interactive intervention of the operators will be limited to the documents that the automated tools are not able to process. Each manually elaborated document becomes then a part of the training set of the ML engine, increasing the probability that next time the automatic systems can handle a similar document.

### *A. Context Aware Processing*

In the SIBDA ECM System, a Context Aware processing model was adopted as a solution to manage the unstructured data complexity. With the aim of managing complexity, the approach was to divide in two stages the problem of selecting the data sought from the amount of unstructured data. In the first stage, a recognition of the context is performed, while in the second stage the process of extracting the searched data is carried out. According to this two-stage decomposition, the complexity of the task of automatic identification of the semantic content is reduced. The information present in the textual content is thus divided into *specific information* to be extracted, and *contextual information*, thus separating the task of analysis into two different tasks and two successive phases. The process separates the phase of extracting the detailed textual data from the phase of recognizing the type of received document, which represents the context of data processing. The same text in one document is therefore interpreted differently, according to the *context*, which is represented by the categorization of the document in which it appears.

To apply this scheme, the adopted solution was to treat the variability of the context as a *classification problem*, describing the context as a class of documents. The categorization is based on a two-level classification tree of document types and subtypes. Currently, the classifier recognizes dozens of document types, each with different subtypes.

The logical architecture of the ECM system is therefore composed of a classifier module that examines the input data and recognizes the context class. A different data extraction module is then invoked according to the assigned textual context. The described scheme is represented in Figure 3.

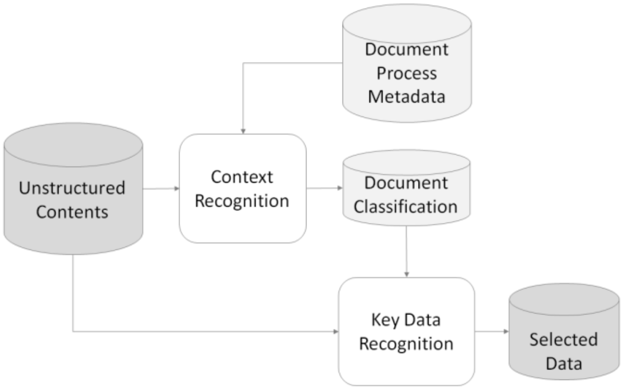

Figure 3. Separating context information from core information

Both the detection of the context and the extraction of the semantic content are supported by a ML engine and by the statistical methods typical of Text Analytics. The contextual information on which the classification is based was then not explicitly defined but the task is assigned to the ML system, which, by carrying out the classification task, recognizes the most likely context and describes it by assigning a class to the text document.

The context recognition module performs a probabilistic evaluation of the document type against a predefined list of document categories. The evaluation is based on a supervised ML approach [12], which requires a preliminary training stage on a set of real customers' documents.

A previous solution, based on the explicit definition of a template for each document type, was tested and later abandoned, since it was not able to manage the variety of texts to be processed with the requested flexibility.

### B. *Process design and adopted techniques*

As for the techniques used in the classification and indexing phases of documents, we distinguish between two categories of documents, namely the documents with *prevalence of text* and the documents with a *prevalence of images*.

In the *classification of text documents*, we apply proprietary ML algorithms based on the Support Vector Machines (SVM) method, working on the weight attributed to a set of keywords. In this case, the system is fed by the text extracted from the training set documents, through which the algorithm learns the list of keywords and related weights, useful for document classification.

For the phase of *data extraction* (i.e., the extraction of the characteristic information of each document), different logics are applied, depending on whether the textual document is structured or unstructured. In case we have *structured documents*, a set of rules is applied, which employs: a) recognition logics based on regular expressions, and b) the search for specific anchor elements in the documents. For *unstructured documents*, a supervised ML algorithm is applied that operates as follows: in the preparation of the training set, the texts are annotated with tags that are affixed to the key information to be extracted. The algorithm, then, learns to recognize the textual patterns related to the occurrence of combinations of words and phrases, and then extracts the relevant text portions from the document.

Regarding the *classification of documents* with prevalence of images, the classifier uses supervised ML algorithms based mainly on neural networks. The training set consists of images on which a pre-processing of the documents is made (in particular a "cropping"), to maximize the effectiveness of the classification.

The subsequent *data extraction* phase starts from the information returned by the classification phase, such as the document type and the position where the text was found. To maximize extraction yield, OCR software is applied, focusing only on specific portions of the document, after an initial pre-processing phase.

To improve the automatic recognition rate, a further analysis stage is applied, using – when possible - a *fuzzy lookup* approach. This approach uses the support data that are often provided by customers, to verify the quality of the extracted data. Through the fuzzy lookup technique, every data extracted from automatic processing is matched against a dictionary of valid words, including the personal data related to the specific customer.

To determine whether to invoke a human operator intervention, three indexes of confidence are computed that provide an estimation of each automated phase reliability (OCR reliability, reliability of classification, reliability of data extraction).

Both the automated mode and the interactive mode processing, share the same process management architecture and process analytics tools, that allow an overall monitoring and a comparative analysis, to evaluate the efficiency of the automated processes.

Upon completion of the classification and extraction phases, the document properties and extracted key data are stored in the databases of the document management system. In order to store both the data extracted from documents and the process data, the selection turned to a document-oriented database with Big-Data capability, which allows for variable data structures according to the type of document. Document-oriented DBs provide structured data types, which allow storing multi-level text structures. For this purpose, the MongoDB server was adopted, which natively uses the JSON (JavaScript Object Notation) format, in an extended variant with binary encoding called BSON.

The complete textual data, i.e., the full text extracted from the document by OCR, are not currently saved, since this operation could possibly violate privacy issues. However, if a customer requires the availability of complete texts, for example to design new value-added services, the documents can be stored as text files on a dedicated file-server.

The overall software infrastructure has a SOAP architecture. This solution was chosen with two objectives: to be highly scalable and to easily support the adding/changing of technological tools, with limited impacts on further processing.

### C. *Experience with the ECM system*

The proposed approach was then prototyped and applied in a defined context, providing for basis the experimentation of different solutions and software tools, leading to the definition of some Proofs of Concept (PoCs). Their purpose was to evaluate the performances, in terms of quality and speed of execution, and to estimate the costs of adoption, both for the implementation effort and for the infrastructures set up, before the passage into production on a set of pilots.

The *success levels* achieved by these techniques, and the levels required for a real production environment vary according to the type of document. However, for structured text documents (such as contracts documentation or privacy consent forms), classification levels considered adequate for a practical use of the system have already been achieved. In fact, by creating training sets with 200/300 *good quality* documents (i.e., readable by OCR software with a low frequency of errors), we reach automatic classification rates higher than 90-95%. These values refer to the classification of several types of documents linked within a single PDF file, where the system identifies the different document types and performs a "splitting". In this type of processing, the documents are specific to each individual processing, so it is necessary to set up dedicated training sets for the different processes.

The identification documents, namely registry documents, IDs, driving licenses and so on, are instead managed as a type of document transversal to the different business processes. Here it was necessary to use a broader training set, reaching the thousand documents for each type (an example of document is the front of an identity card), to achieve even in this case automatic classification percentages around the 90-95%.

As regards the phase of data extraction (indexing) from the identification documents, the percentages achieved are lower, due to the poor quality of the original digitized document image and the extreme variability of the texts. Using the approach described above, we have achieved in this case an automatic recognition rate of around 75-80%.

The transition of the pilot processes into production will allow us to test the competitiveness of the current architecture and methods, so as to support the selection of the best techniques to implement, gradually extending their use to an ever-increasing number of document processes.

The next steps we are taking to further develop the system are currently in two directions: the experimentation of techniques of semantic analysis of documents and the identification through neural networks of the signatures in the documents.

## V. CONCLUDING REMARKS

This paper presented a context-aware architecture for a BD process. Currently, the experimentation is moving towards the production phase and the first operations on an actual document flows are under implementation by Microdata. At the architectural level, we believe that this project shows that under a context-aware approach it is possible to tackle some tasks of analyzing highly variable and unstructured data through classification techniques, thus reducing the complexity of the work required at the core of the text analysis processes.

Future work includes, among other issues, the study of privacy related to managing ECM data.

## REFERENCES

[1] S.K. Shivakumar. "Enterprise content and search management for building digital platforms". John Wiley & Sons, 2016.

[2] S.E. Bibri, and J. Krogstie. "ICT of the new wave of computing for sustainable urban forms: their Big Data and context-aware augmented typologies and design concepts". *Sustainable cities and society*, 32, 2017, pp.449-474.

[3] A.G. De Prado, G. Ortiz, and J.Boubeta-Puig,. "CARED-SOA: A context-aware event-driven service-oriented Architecture", IEEE Access, 5, (2017), pp.4646-4663

[4] O.B. Sezer, D.Erdogan, and Ahmet Murat Ozbayoglu. "Context-aware computing, learning, and BIG DATA in Internet of Things: a survey." *IEEE Internet of Things Journal* 5.1 (2018), pp.1-27.

[5] S.E. Bibri, and J. Krogstie. "The core enabling technologies of Big Data and context-aware computing for smart sustainable cities: a review and synthesis." *Journal of Big Data* 4.1 (2017): 38.

[6] R. Sandhu, K.Jaspree, and V. Thapar. "An effective framework for finding similar cases of dengue from audio and text data using domain thesaurus and case base reasoning." *Enterprise Information Systems* 12.2 (2018): 155-172.

[7] K.P. Subbu and A. V. Vasilakos. "Big Data for context aware computing–perspectives and challenges." *Big Data Research, 10* (2017): pp.33-43.

[8] M.G.Fugini, and J. Finocchi, "Innovative Big Data Analytics: A System for Document Management", *IEEE WETICE W2T'18 Conf*., Paris, June 2018.

[9] “Big Data: fast and smart”, Research Report, Observatories at Politecnico di Milano, Observatory on Big Data Analytics & Business Intelligence Observatory, 2018. Available: http://www.osservatori.net

[10] “Big Data is now: tomorrow is too late”, Research Report Observatories at Politecnico di Milano, Observatory on Big Data Analytics & Business Intelligence, February 2017. Available http://www.osservatori.net

[11] “Corporate Entrepreneurship e Open Innovation: innovare con un occhio alle startup!”, Research Report, Observatories at Politecnico di Milano, Startup Intelligence Observatory, 2017. Available http://www.osservatori.net

[12] V.Bijalwan, et al. "Machine learning approach for text and document mining." arXiv preprint arXiv:1406.1580 (2014).