1 Publishable summary

Overall Project Context and Objectives

Many collaboration and decision making settings are nowadays associated with huge, ever-increasing amounts of multiple types of data, obtained from diverse sources, which often have a low signal-to-noise ratio for addressing the problem at hand. These data may also vary in terms of subjectivity and importance, ranging from individual opinions and estimations to broadly accepted practices and trustable measurements and scientific results. Additional problems start when we want to consider and exploit accumulated volumes of data, which may have been collected over a few weeks or months, and meaningfully analyze them towards making a decision. Admittedly, when things get complex, we need to identify, understand and exploit data patterns; we need to aggregate appropriate volumes of data from multiple sources, and then mine them for insights that would never emerge from manual inspection or analysis of any single data source. In these settings, "big data" analytics technology currently receives much criticism, in that it does not provide proper insight into what the data means. To make sense of big data and come with discoveries that help improve decision making in practical contexts, human intelligence should be also exploited. We need to provide the appropriate ways to nurture and capture this human intelligence in order to extract the necessary insights and improve the way machines deal with complex situations.

Taking the above issues into account, the Dicode project aims at facilitating and augmenting collaboration and decision making in data-intensive and cognitively-complex settings. To do so, whenever appropriate, it builds on prominent high-performance computing paradigms and large data processing technologies to meaningfully search, analyze and aggregate data existing in diverse, extremely large, and rapidly evolving sources. At the same time, particular emphasis is given to the deepening of our insights about the proper exploitation of big data, as well as to collaboration and sense making support issues. Building on current advancements, the solution foreseen in the Dicode project brings together the reasoning capabilities of both the machine and the humans. It can be viewed as an innovative "workbench" incorporating and orchestrating a set of interoperable services that reduce the data-intensiveness and complexity overload at critical decision points to a manageable level, thus permitting stakeholders to be more productive and effective in their work practices. Services that are developed and integrated in the context of the Dicode project are released under an open source license.

The achievements of the Dicode project are being validated through three use cases:

- ***Clinico-Genomic Research Assimilator.*** The need to collaboratively explore, evaluate, disseminate and diffuse relative scientific findings and results is more than profound today. Towards this objective, Dicode envisages to plan an integrated clinico-genomic (tacit) knowledge discovery and decision making use case that targets the identification and validation of predictive clinico-genomic models and biomarkers.
- ***Trial of Clinical Treatment Effects.*** The goal of this case (which has been expanded in the second year of the project to cover broader clinical trials, not just for Rheumatoid Arthritis) is to facilitate the process of making clinical decisions in drug trials by combining datasets from patient results (blood tests, physical examinations) and the different scan modalities (X-Ray, Static and Dynamic MRI scan images) to reveal the effectiveness of a drug within a trial.
- ***Opinion Mining from unstructured Web 2.0 data.*** Through this case, we aim to validate the Dicode services for the automatic analyses of the voluminous amount of unstructured information existing on the Web, especially in the highly dynamic social media space. Data for this case will be primarily obtained from spidering the most popular social Web sites making use of APIs from various Web 2.0 platforms.

Work Performed and Main Results Achieved so far

The work performed in Dicode follows an evolutionary approach, where: (i) both stakeholders and technology developers are being actively engaged in the specification, design and evaluation of the foreseen technological solutions; (ii) innovative services are being developed incrementally to ensure that end users can experiment with the Dicode services as early as possible; (iii) user requirements are being refined through testing and trials, involving users from the three use cases.

Work carried out so far (during the first two years of the project) mainly concerns the following activities:

- Analysis of current practices and user requirements as far as data-intensive and cognitively-complex collaboration and decision making is concerned;
- Specification of the overall Dicode approach;
- Setup and upgrade of the Dicode infrastructure;
- Development and integration of Dicode services according to the foreseen research objectives;
- Specification of a comprehensive Dicode evaluation framework and first round of evaluation of the Dicode services;
- Dissemination and exploitation of the project's activities and outcomes.

Analysis of Current Practices and User Requirements

Requirements have been collected from the three use cases of the project. Our aim was to understand current practices, focusing on their data intensive decision making and collaborative activities, as well as on the characteristics of data and their usage. A Dicode specific requirement elicitation strategy was designed and deployed to tackle the seemingly diverse use cases. A rich set of usage scenarios was produced and elaborated, involving both end users and technology developers. These scenarios helped significantly towards establishing a consensus of the role of the envisioned suite of Dicode services. Thoroughly considering the feedback from the first evaluation round of Dicode services across the project's use cases, an analysis of the lessons learned was documented and services' specifications were revised to inform the next iteration of development. A deeper understanding of the use cases' differences and similarities, as well as of their potential to explore the full range of Dicode services was achieved.

Specification of the Overall Dicode Approach

A high level Dicode work methodology - addressing issues such as the meaningful involvement of stakeholders, iterative cycles of service development, agility of development efforts, as well as the exploration and consolidation of ideas for innovative work practices - was elaborated and agreed in the early stages of the project. Moreover, an initial conceptual architecture for the Dicode platform was developed to guide the brainstorming sessions for potential solutions with all partners. Existing generic (individual and collaborative) sense-making frameworks were adopted as a basis for identifying a common strand across the three use cases in terms of tackling data-intensive and cognitively-complex collaboration and decision making activities. This showed potential in guiding the development of innovative work practices together with the foreseen Dicode services. The overall Dicode approach was updated after taking into account feedback from the first evaluation round of Dicode services.

Setup of the Dicode Infrastructure

The basic infrastructure on in-house computer clusters for running large scale data mining experiments and testing prototype implementations, as well as data collections for benchmarking based on textual data and structured data were set and upgraded during the project's evolution. In addition, standards and guidelines for the development of Dicode services - aiming at ensuring interoperability between the services to be developed and reusability of them through diverse scenarios of use – were defined. Through this work, partners agreed on technical requirements, as well as on development technologies and supporting tools to handle issues such as management of source code repository, software building management and bug tracking. Issues around the conceptual integration of the foreseen Dicode services were thoroughly elaborated.

Development and Integration of Dicode Services

According to the foreseen workplan, advanced operational versions of the Dicode Data Mining Services, the Dicode Collaboration Support Services, and the Dicode Decision Making Support Services have been produced and integrated (these versions took into account the feedback collected from the first evaluation round of the project). The proposed Data Mining Framework, which is considered as an instance of the Service-Oriented Architecture, takes into account the particularities of the Dicode project, and specifies the list of services to be offered in Dicode. The Dicode Data Mining Services developed so far concern text mining, designating text mining (dealing with different facets of opinion mining), subgroup discovery, recommendation and similarity learning. The Dicode Collaboration Support Services exploit the reasoning abilities of humans to facilitate sense-making; services developed so far offer innovative virtual workspaces that support collaboration towards sense-making in data intensive settings, as well as indexing and searching (full-text and meta-data search) of standard documents existing in these workspaces. Finally, the Dicode Decision Making Support Services aim to meaningfully build machine-interpretable knowledge in order to actively support various decision making tasks; services developed so far concern data mining for community modelling and the intelligent support of stakeholders in decision making activities by enabling alternative reasoning mechanisms. The integration of Dicode services is performed within the Dicode Workbench (Figure 1), a web application offering a series of functionalities to instantiate the Dicode solution for each use case and enable innovative work methodologies.

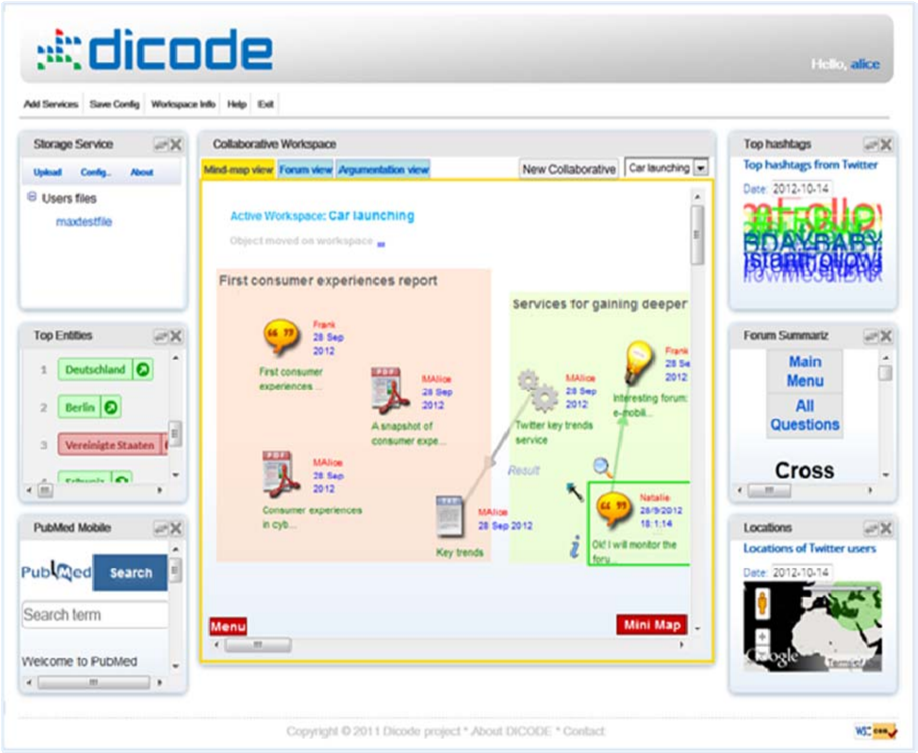

Figure 1 : An instance of the Dicode Workbench

Specification of a Dicode Evaluation Framework and First Evaluation Round

A comprehensive evaluation framework was specified in the first year of the project, identifying a broad range of aspects through which the foreseen Dicode services will be evaluated. Its first part concerns the identification of Dicode Key Success Indicators, which ensure that the overall Dicode objectives will be met. The second part is devoted to indicators such as quality of services offered, improvement of productivity and creativity, Dicode solutions' usefulness and ease-of-use, as well as adaptability, accessibility and acceptability of the Dicode services. The first round of evaluation of the Dicode services through the project's use cases was performed in the second year of the project. Properly formulated metrics and questionnaires were employed to analyse the feedback received. Each service specific questionnaire was accompanied with help files provided by technical partners.

Dissemination and Exploitation Activities

A comprehensive exploitation and dissemination plan has been produced, ensuring the impact and sustainability of the Dicode outcomes. Initial dissemination and exploitation activities include the development of a corporate identity of the project, the set-up of a web portal, and initial public relations efforts. A significant number of publications have resulted out of joint work among consortium members. These publications appear in international scientific journals and proceedings of international peer-reviewed scientific conferences and workshops. Presentations of project-related work were also given in some of the top conferences on open source high scalability technologies. Moreover, Dicode organized three scientific workshops, one in the context of the world leading conference on collaboration support (CSCW 2012), another in the context of the best European conference on machine learning and knowledge discovery (ECML-PKDD 2012), and a third one at the leading international conference on knowledge engineering and knowledge management (EKAW2012). A series of activities with respect to the availability of Dicode software for public use, the continuous support to active and interested workgroups that use the Dicode services and the maintenance of close contacts to the industry has also taken place. While dissemination towards research communities was the main focus during the first year of the project, exploitation became more important in the second year. Dicode partners have already reported several success stories concerning exploitation of Dicode results, development of strategic partnerships with industry and co-operation with other EU projects.

Expected Final Results and their Potential Impact and Use

The final results of the Dicode project are expected to advance the state-of-the-art in approaches on (i) the proper exploitation of big data (the "big data fallacy" issue) and the integrated consideration of data mining and sense-making issues, (ii) recommender systems, with respect to recommendations in heterogeneous, multi-faceted data and the identification of hidden links in complex data types, (iii) understanding text to drastically reduce the annotation effort for extracting relations, (iv) opinion mining by considering opinion statements as n-ary relations and apply the highly scalable methodology implemented for their recognition, (v) Web 2.0 collaboration support tools in terms of interoperability with third party tools and integration of appropriate reasoning services, and (vi) decision making support applications, by integrating knowledge management and decision making features as well as by building on the synergy of human and machine argumentation-based reasoning.

The foreseen advancements will ultimately shape innovative work methodologies for dealing with the problems of information overload and cognitive complexity in diverse collaboration and decision making contexts. Both individual and collaborative sense making will be augmented through the meaningful exploitation of prominent data processing and data analysis technologies. The foreseen solution will be user-friendly and built on the synergy of human and machine intelligence. It will mask the overall complexity of the underlying issues, thus allowing stakeholders to easily interact with large and complex data, providing them with meaningful recommendations upon which they can base their decisions and actions. Moreover, machine-tractable knowledge concerning the full lifecycle of collaboration and decision making will be accumulated and maintained. Consequently, the foreseen solution will augment the productivity of stakeholders.

Adopting open standards, and in accordance with EU's recent initiatives on Open Systems and Data, the Dicode project has the potential of forming a rich ecology of domain specific and nonspecific extensions. The foreseen solution will allow for external data service providers to supply information, as well as for external developers to supply additional modules and applications, which are tailored to evolving market conditions. Finally, it will enable diverse public and private entities to aggregate, structure, semantically enrich and analyse vast amounts of information. This turns the problem of information overload into a benefit of structured data, which can be used as the basis for decisions of better quality. Simply put, the foreseen solution aims at turning information growth into economic growth.

Project Public Website

http://dicode-project.eu

Contact Details

Nikos Karacapilidis
Computer Technology Institute & Press "Diophantus"
26504 Rio Patras, Greece
E-mail: info@dicode-project.eu, karacap@cti.gr

List of Partners

- Computer Technology Institute & Press "Diophantus" (project coordinator)
- University of Leeds
- Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V.
- Universidad Politécnica de Madrid
- neofonie Gmbh
- Image Analysis Ltd
- Biomedical Research Foundation - Academy of Athens
- Publicis Frankfurt Zweigniederlassung der PWW GmbH