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 big 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 find patterns that really help organizations make better business decisions, 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, it exploits and 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 collaboration support and sense making issues, aiming to satisfy the needs of end users towards making better decisions. Building on current advancements, the solution foreseen in the Dicode project brings together the reasoning capabilities of both the machine and the humans (Figure 1). 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.

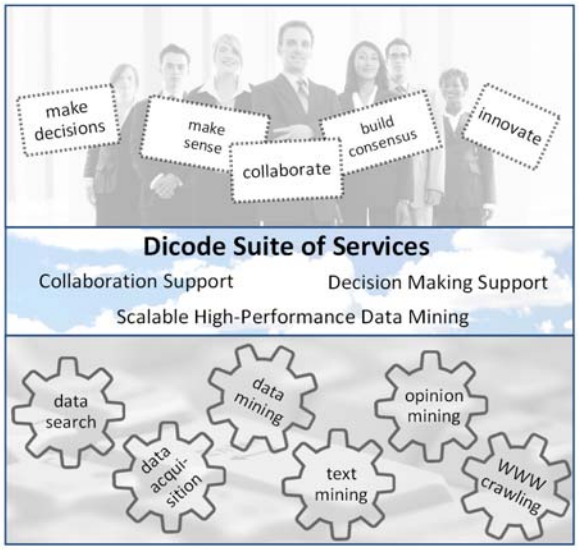

Figure 1: *The Dicode services build on the synergy of machine and human reasoning.*

The achievements of the Dicode project are being validated through three use cases. These were chosen to test the transferability of Dicode solutions in different collaboration and decision making settings, associated with diverse types of data and data sources, thus covering the full range of the foreseen solution's features and functionalities. These cases concern:

- ***Clinico-Genomic Research Assimilator.*** This case will demonstrate how Dicode can support clinico-genomic scientific research in the current post-genomic era. 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. The use case is founded on the seamless integration of both heterogeneous clinico-genomic data sources and advanced analytical techniques provided by Dicode.

- ***Trials of Rheumatoid Arthritis (RA) Treatment and clinical trials.*** This case will benefit from Dicode services to deliver pertinent information to communities of radiologists, doctors and patients in the domain of randomised clinical trials of potential new drugs. This was originally confined to RA but is now extended to drugs in all diseases. Drug trials are carried out by an academic research establishment on behalf of a pharmaceutical company. Each trial will evaluate the effectiveness of treatment for disease by analysing the condition in patients receiving the drug and those receiving a placebo. Dicode services will be used to enable an affective and collaborative way of working towards decision making by various individuals involved (Radiographers, Radiologists, Clinicians, etc.). Dicode services will enable data mining, clustering and classification, as well as collaboration tools for doctors working on multiple sites.
- ***Opinion Mining from unstructured Web 2.0 data.*** It is paramount today that companies know what is being said about their services or products on the Web, especially in the highly dynamic social media space. With the current tools, finding who and what is being said on the social web is literally searching for a needle in the haystack of unstructured information. Through this case, we aim to validate the Dicode services for the automatic analyses of this voluminous amount of unstructured information. 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, such as micro-blogging platforms (Twitter), social network platforms (Facebook), as well as relevant blogs and forums.

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 at early stages; (iii) user requirements are being refined through testing and trials (involving users from the three use cases).

Work carried out during the first year 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 of the Dicode infrastructure;
- Development of Dicode services according to the foreseen research objectives;
- Specification of a comprehensive Dicode evaluation framework;
- Development and deployment of a plan for the dissemination of the project's activities and outcomes.

Analysis of Current Practices and User Requirements

Requirements were 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. This resulted in a deeper understanding of the common characteristics across all three use cases, as well as case-specific requirements and vision. 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. In addition, state-of-the-art technologies and technical challenges ahead were explored and documented. State-of-the-art was reviewed in the areas of scalable high-performance data mining, data mining towards sense-making of real-world multi-faceted data, collaboration and decision making support, and integration technologies. Their relevance and applicability were assessed with respect to the capturing of tractable information, the delivering of pertinent information, the overall collaboration and decision making support.

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. Moreover, an initial conceptual architecture for the Dicode platform was developed to guide the brainstorming sessions for potential solutions with all partners. Existing generic 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 practice together with the foreseen Dicode services. In addition, functional specifications were elaborated and documented for ten groups of services identified: Data Mining Services, Web-Data Acquisition Services, Data Mining Services on Textual Data, Collaboration and Decision Making Services, Semantic Services (Ontology Querying and Content Annotation), Community Modeling and User Profiling, Data Acquisition Services for Heterogeneous Resources, Data Pre-processing Services, Data Analysis Services on Heterogeneous Datasets, and Integration Services.

Setup of the Dicode Infrastructure

The basic infrastructure (i.e. Hadoop, Mahout) 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 (i.e. Apache Software Foundation mailing lists, Wikipedia) and structured data (i.e. Yahoo collection) were set. Experimentation with the existing Mahout machine learning algorithms took place. 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, including: (i) the specification of an ontological framework for the capture and representation of diverse stakeholder perspectives to augment collaboration and decision making was outlined, (ii) the development of an intuitive ontology engineering tool extended with intelligent features to provide semantic feedback, (iii) development of the first version of the Dicode Ontology, and (iv) development of services (and an associated prototype) which semantically tag and link outputs from human decision/sense making processes.

Development of Dicode Services

According to the foreseen workplan, initial versions of the Dicode Data Mining Framework, the Dicode Data Mining Services, the Dicode Collaboration Support Services, and the Dicode Decision Making Support Services were produced. 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 appropriate reasoning mechanisms. Work on the integration of operational versions of the above services has also started to be integrated, thus contributing to the instantiation of the Dicode suite of services for each use case.

Specification of a Dicode Evaluation Framework

A comprehensive framework was specified identifying a broad range of aspects through which the foreseen Dicode services could 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), which can be included in the evaluation of the Dicode services. A list of relevant data collection and analysis instruments has been also composed, which can provide guidance for the planning of the forthcoming evaluation trials for each service.

Dissemination Activities

During the first year of the project, dissemination and exploitation activities included the development of a corporate identity of the project, the set-up of a web portal, and initial public relations efforts. In addition, a number of publications resulted out of joint work among consortium members. These publications appeared in international scientific journals and proceedings of international peer-reviewed scientific conferences and workshops. Presentations of Dicode related work were also given in some of the top conferences on open source high scalability technologies. Moreover, Dicode people co-organized the International Workshop on Adaptive Support for Team Collaboration (held in conjunction with UMAP 2011), as well as organized the "Semantic/NLP Hackathon" at "Berlin Buzzwords – The Conference of High Scalability". Our proposal to organize the 1st Dicode workshop, entitled "Mastering Data-Intensive Collaboration through the Synergy of Human and Machine Reasoning", in the context of the world leading conference on collaboration support - CSCW 2012 - has been accepted. Finally, a series of activities with respect to the availability of software prototypes 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 took place. Strategic partnerships with industry sought so far concern marketing and medical field organizations.

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) recommender systems with respect to recommendations in heterogeneous, multi-faceted data and the identification of hidden links between different pieces of information, (ii) the identification of relevant links in complex data types, heterogeneous multi-faceted data, and very large data sets, (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 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. 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, the Dicode project has the potential of forming a rich ecology of domain specific and non-specific 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, and Publicis Frankfurt Zweigniederlassung der PWW GmbH