Mastering Data-Intensive Collaboration and Decision Making

FP7 - Information and Communication Technologies

Grant Agreement no: 257184
Collaborative Project
Project start: 1 September 2010, Duration: 36 months

D2.3 – The Dicode Approach Revisited

Due date of deliverable:	31 August 2012
Actual submission date:	31 August 2012
Lead Partner:	UOL
Contributing Partners:	UOL, BRF, IMA, PUB, CTI, FHG, NEO, UPM

Nature: ☒ Report ☐ Prototype ☐ Demonstrator ☐ Other

Dissemination Level:

☒	PU :	Public
☐	PP :	Restricted to other programme participants (including the Commission Services)
☐	RE :	Restricted to a group specified by the consortium (including the Commission Services)
☐	CO :	Confidential, only for members of the consortium (including the Commission Services)

Keyword List: Requirement analysis, requirement elicitation, use case, architecture, methodology, functional specification

The Dicode project (dicode-project.eu) is funded by the European Commission, Information Society and Media Directorate General, under the FP7 Cooperation programme (ICT/SO 4.3: Intelligent Information Management).

The Dicode Consortium

Research Academic Computer Technology Institute (CTI) (coordinator), Greece

University of Leeds(UOL), UK

Fraunhofer-Gesellschaft zur Foerderung der angewandten Forschung e.V. (FHG), Germany

Universidad Politecnica De Madrid(UPM), Spain

Neofonie GmbH(NEO), Germany

Image Analysis Limited(IMA), UK

Biomedical Research Foundation, Academy of Athens(BRF), Greece

Publicis Frankfurt GmbH(PUB), Germany

Document history			
Version	Date	Status	Modifications made by
1	05-07-2012	First Draft	Fan Yang-Turner
2	12-07-2012	Second Draft (with comments from Vania Dimitrova and Lydia Lau)	Fan Yang-Turner
3	18-07-2012	Third Draft (circulated to internal reviewers and Dicode partners for feedback)	Fan Yang-Turner, Lydia Lau
4	17-08-2012	Fourth Draft (feedback incorporated, sent to SC for approval)	Fan Yang-Turner, Lydia Lau
5	31-08-2012	Final version (approved by SC, sent to the Project Officer)	Fan Yang-Turner, Lydia Lau

Deliverable managers

- Lydia Lau (UOL)
- Fan Yang-Turner (UOL)

List of Contributors

- Georgia Tsiliki (BRF, Use Case 1)
- Martin Hinton (IMA, Use Case 2)
- Ralf Loeffler (PUB, Use Case 3)
- Manolis Tzagarakis (CTI)
- Spyros Christodoulou (CTI)
- Doris Maassen (NEO)
- Jörg Kindermann (FHG)
- Natalja Friesen (FHG)
- Guillermo de la Calle Velasco (UPM)
- Vania Dimitrova (UOL)
- Dhavalkumar Thakker (UOL)

List of Evaluators

- Nikos Karacapilidis (CTI)
- Guillermo de la Calle Velasco (UPM)

Executive Summary

This document is the final deliverable of WP2 which is dedicated to acquire a deep understanding of user requirements and system characteristics which are fundamental to the Dicode concept and approach. The Dicode approach (previous deliverable, D2.2) is revisited and updated by acquisition of new knowledge (individual and collaborative sensemaking activities in the three use cases) and by reflection on the achievements to date (the development of Dicode services and user trial results as described in D6.2.1, D6.3.1 and D6.4.1). This updated Dicode approach includes the methodology, the conceptual architecture and usage scenarios of each use case.

The Dicode methodology has brought use case partners and technical partners together and enables the development of Dicode services in an incremental manner. This methodology reinforces the equal partnership between end users and technical partners. This mutual and co-evolutionary understanding is crucial in the exploration of an innovative solution for fostering synergy between human and machine reasoning.

Big data analytics has been identified as the global conceptual space for the Dicode project. All users from the three use cases are professional knowledge workers who require intelligent support from the Dicode platform to make better sense of the data either on their own or collaboratively with other professionals.

Underpinned by sensemaking models, the commonalities and differences of three use cases have been analysed. Dicode use cases are common in performing the bottom-up information processing according to Pirolli and Card's framework [8] (search and filter, read and extract, schematize, build case, and tell story). The differences among the three use cases are the types and sources of the data being processed and transformed, either structured or unstructured and drawing from three different domains: biomedical research, clinical trial and social media monitoring.

These commonalities and differences have informed the generic conceptual architecture for Dicode, which supports both individual sensemaking and collaborative sensemaking through three types of services: data-centric services, collaboration-centric services and integration-centric services. Data-centric services exploit large-data processing technology to meaningfully search, analyze and aggregate data from heterogeneous data sources in order to improve the processes of individual sensemaking. Collaboration-centric services support users and their interaction by capturing and sharing resources, opinions, arguments and comments among participants, and to facilitate the collective understanding of the issues related to data analysis. Integration-centric services support data-centric services and collaboration centric-services as well as the users with the aim to ensure and facilitate the seamless – from both the conceptual and the technical point of view – interoperability and integration of the independent services developed.

The generic conceptual architecture has been instantiated to reflect the adoption of Dicode services by each use case according to their unique characteristics of the associated data and processes. The usage scenarios have demonstrated how these adopted Dicode services help users in their intelligent analysis individually and collaboratively. This generic architecture and its instantiations have shown that Dicode solution can be applied to different domains and can be evolved with increasing understanding and contributions towards big data analytics.

Table of Contents

Executive Summary **4**

1 Introduction **6**
1.1 Context 6
1.2 Objectives and structure 6

2 The Dicode Methodology **6**

3 Conceptual Underpinning: Sensemaking Frameworks **8**
3.1 Individual Sensemaking Framework 9
3.2 Collaborative Sensemaking Framework 12
3.3 Generic Conceptual Architecture for Dicode 14

4 Use Case 1: Clinico-Genomic Research Assimilator (CGRA) **16**
4.1 Instantiated Conceptual Architecture of Use Case 1 16
4.2 Dicode Services and Usage Scenario of Use Case 1 17

5 Use case 2: Trial of Clinical Treatment Effects (TCTE) **18**
5.1 Instantiated Conceptual Architecture of Use Case 2 18
5.2 Dicode Services and Usage Scenario of Use Case 2 19

6 Use case 3 - Opinion Mining from Unstructured Web 2.0 Data (OPUWD) **20**
6.1 Instantiated Conceptual Architecture of Use Case 3 20
6.2 Dicode Services and Usage Scenario of Use Case 3 21

7 Conclusion **22**

References **23**

1 Introduction

1.1 Context

This document is the final deliverable of WP2, which is dedicated to acquire a deep understanding of user requirements and system characteristics which are fundamental to the Dicode concept and approach. The Dicode approach (previous deliverable, D2.2) is revisited and updated by acquisition of new knowledge (individual and collaborative sensemaking activities in three use cases) and by reflection on the achievements to date (the development of Dicode services and user trial results as described in D6.2.1, D6.3.1 and D6.4.1).

The approach taken by the project towards mastering data-intensive collaboration and decision making, i.e. the **Dicode Approach**, encompasses three themes:
- Methodology,
- Conceptual underpinning, and
- Architectural framework.

The commonalities and differences of three use cases inform and drive the development of the Dicode architectural framework and services.

1.2 Objectives and structure

This document firstly presents the updated **Dicode Methodology,** which reinforces the equal partnership between end users and technical partners. This mutual understanding is crucial in the exploration of an innovative solution for fostering synergy between human and machine reasoning [11].

Big data analytics is the global conceptual space for the Dicode project. All users from the three use cases are professional knowledge workers who require intelligent support from the Dicode platform to make better sense of the data either on their own or collaboratively with other professionals. Section 3 dissects the commonalities and differences amongst the three use cases, using the analytical lenses from theories for **individual and collective sensemaking**. The outcome of this understanding informs the refinement of the Dicode conceptual architecture.

The Dicode conceptual architecture aims to provide a common framework to guide the direction of investigation and experimentation in the project. When it comes to extending our work for the solution space, **instantiation of the architecture for each use case** is necessary to deepen our understanding in the likely impact of unique characteristics of the associated data and processes. In other words, study of differences may lead a more effective ‘plug and play’ generic Dicode platform in the future. Sections 4, 5 and 6 describe the sensemaking activities and the instantiated architecture of the Dicode use cases.

2 The Dicode Methodology

The main purpose of the Dicode Methodology is to synchronise the range of activities undertaken by the consortium partners to address the challenge of fostering synergy between human and machine intelligence in data-intensive environments.

Figure 1 presents the Dicode methodology as implemented, which has been slightly modified from the initial version in D2.2 which positioned stakeholders more prominently between requirements and evaluation stages. In this update, the stake holders are placed in a more central role in the methodology to emphasize their constant involvement during the three stages of project, and in particular their timely input into the design and development of services: The three stages, however, remain unchanged as follows:

1. User and architectural requirements elicitation;
2. Iterative development of Dicode services (and their integration);
3. Service review and/or use case evaluation.

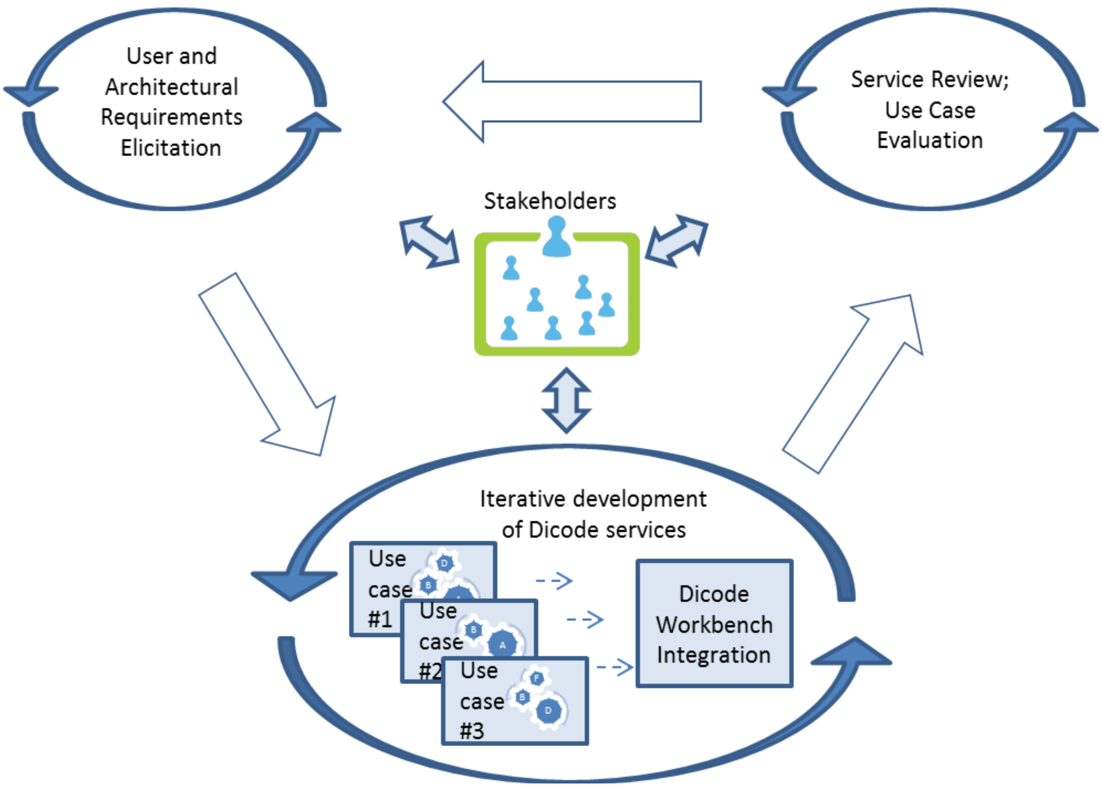

Figure 1. The Dicode Methodology

This three-stage methodology iterates twice in the project to enable the consolidation of user and architectural requirements during mid-project in response to feedback.

Iteration I (month 1 to month 18)
During iteration I, initial user and architectural requirements have been collected and services have been designed to meet requirements of individual use cases.

Integration started at month 12 at two levels: (i) technical integration of services into the Dicode workbench, and (ii) user trials when each use case partner evaluated the delivered services with the view of integrating them into their work practice.

Iteration II (month 19 to month 36)
Iteration II started with reflecting on the lessons learned from Iteration I. New services were developed to meet new requirements; existing services were extended for reuse in another use case. The conceptual and technical architecture for integration were consolidated. This iteration will end with completion of planned services and the final round of user trials. The activities of development, evaluation and user analysis have been constantly updated through the Dicode wiki web site[1].

[1] https://wiki.dicode-project.eu/dashboard.action

3 Conceptual Underpinning: Sensemaking Frameworks

At the outset of the Dicode project, three use cases were selected as they presented the following challenges in collaboration and decision making in data-intensive and cognitive-complex environment:

- Users experience the problem of information overload;
- Users conduct cognitively intensive analysis and interpretation of data; and
- Discussions, shared interpretation and rationale are needed among specialists.

These common issues are related to a newly forming area for research, namely **big data analytics**.

"*Fundamentally, big data analytics is a workflow that distils terabytes of low-value data (e.g., every tweet) down to, in some cases, a single bit of high-value data (Should Company X acquire Company Y? Can we reject the null hypothesis?). (p. 50) [1]*"

Towards supporting big data analytics, we started with identifying the context of Dicode's three use cases as summarised in Table I. From a high level perspective, these use cases are from different domains; the users have different expertise and use different analytics tools. They deal with different data from different data sources, with stakeholders making different decisions for different purpose in their work. However, all of them are dealing with **intelligent analysis** to transform **input data** into **knowledge product** in order to see the "big picture" from a large collection of information.

	Use Case 1	**Use Case 2**	**Use Case 3**
Application domain	Biomedical Research	Clinical Trials	Social Media Monitoring
Users	Biologists Bio-informaticians	Radiologists; Clinicians	Marketing Analysts; Social Media Analysts
Expertise of users	Biologists Bio-informaticians	Radiography; Clinical Medicine	Marketing; Communications
Analytics tools	Data collection, manipulation and analysis tools (such as R)	Image Analysis tools (such as Dynamika)	Social Media Monitoring Tools
Access of data sources	Public and Private (to research lab)	Private (to the clinical trials)	Public
Input data	Genomic, Transcriptomic Data (DNA sequence data molecular pathways, etc.)	Clinical Data (X-Ray, Static and Dynamic MRI scan modalities)	Social Media Data (Blogs, forums, tweets etc.)
Activities of Intelligent Analysis	Interpreting Result; Planning future research	Concluding the trial result; Planning further trials	Formulating strategy; Planning marketing campaign
Knowledge product	Scientific Findings; Insights for experimental work (e.g. drug design)	Clinical decision; Drug trial result	Strategy for social media engagement

Table I. Different contexts of the three Dicode Use Cases

To better understand the three use cases in terms of **intelligent analysis**, we have chosen an individual sensemaking framework and a collaborative sensemaking framework to guide our investigation. In particular, we are seeking answers to the following research questions:

> *Q1: What are the **commonalities** and **differences** among these three use cases in terms of intelligent analysis?*
>
> *Q2: Can we design a **generic framework** to support those commonalities and differences as exhibited by the three use cases?*

In the following subsections, we present the adopted sensemaking frameworks and derive the commonalities and differences of three use cases in terms of intelligent analysis. We then present Dicode's general conceptual architecture to support those commonalities and differences.

3.1 Individual Sensemaking Framework

In their study, Pirolli and Card [7] found that the basis of an analyst's skill is to quickly organize the flood of incoming information and present his/her analysis in reports. The process of creating a representation of a collection of information that allows the analyst to perceive structure, form and content within a given collection is defined as *sensemaking*.

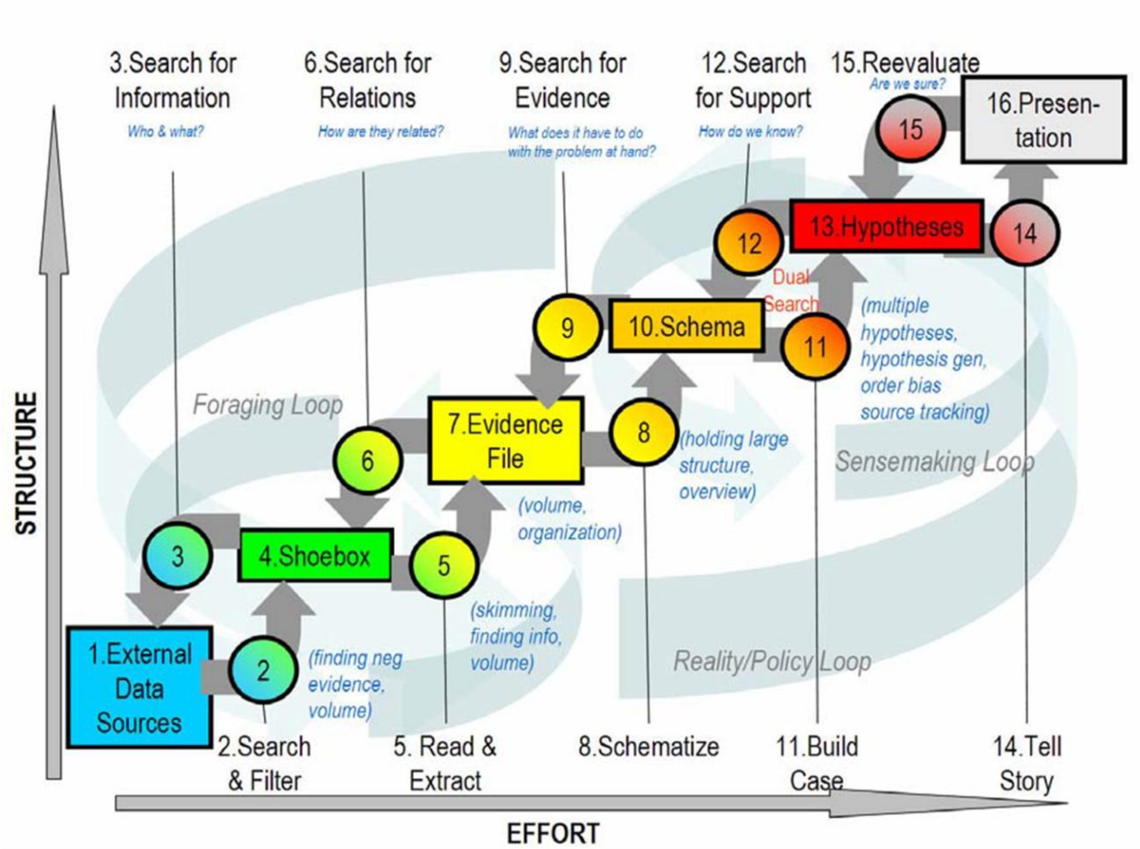

Figure 2. Individual Sensemaking Framework [8]

Sensemaking is an integral part of analysis and it has been modelled in different domains. Russell et al. [9] studied cost structure of sensemaking and modelled sensemaking as cyclic processes of searching for external representations and encoding information into these representations to reduce the cost of tasks to be performed. Pirolli and Card [8] looked into information transformation, claimed that sensemaking is a process of transformation of information into a knowledge product described in a sensemaking process diagram (Figure 2). This diagram summarizes how an analyst comes up with new information. The sequence of rectangular boxes represents an approximate data flow. The circles represent the process flow. The processes and data are arranged by degree of effort and degree of information structure. This is a process with lots of backward loops and seems to have one set of activities that cycle around finding information and another that cycles around making sense of the information, with plenty of interaction between these. According to the framework, the overall information processing can be driven by bottom-up processes (from data to theory) or top-down (from theory to data) and their analysis suggested that top-down process (process 2, 5, 8, 11, 14 in the diagram) and bottom-up processes (process 15, 12, 9, 6, 3) are invoked in an opportunistic mix.

Applying Pirolli and Card's framework to answer our first research question, the Dicode use cases are common in performing the bottom-up information processing (listed below). The differences among the three use cases are the data being processed and transformed. Common aspects of the Dicode use cases include:

- **Search and filter**
  External data sources, such as Pubmed, the Web, or classified databases, provide a repository that is searched (queried) by the analyst. Results of those searches are filtered (judged) for relevance. An analyst filters message traffic or does active search, collecting relevant documents into some kind of storage (the "shoebox" in the diagram) for further processing.

- **Read and extract**
  Collections of shoebox evidence are read to extract nuggets of evidence that may be used to draw inferences, or support or disconfirm theory. Relevant snippets from this store and related low level inferences are placed in evidence files. Evidence extracted at this stage may trigger new hypotheses and searches.

- **Schematize**
  At this point the information may be represented in some schematic way. Given a lack of easily-used tools, this may be in the mind of the analyst, informal, a simple marshalling, or an elaborated computer-based method, for example, a timeline visualisation to coordinate many events. Evidence may be organized into small-scale stories about typical topics or in answer to typical questions (e.g., who, what, when, where, why, how) that are used to organize raw evidence.

- **Build case**
  A theory or case is built by additional marshalling of evidence to support or disprove hypotheses.

- **Tell story**
  A presentation or publication of a case is made to some audience (client).

In Figures 3, 4, 5, examples of these common processes have been illustrated in the context of each use case. The overall information processing of three use cases are mainly driven by bottom-up process (data to theory). However, this does not prevent a number of backward loops and interactions between the steps. For example, in Dicode use case 1, the analysis may be driven by "initial research questions" to find relevant datasets and then further analysis is conducted to derive patterns from those relevant datasets.

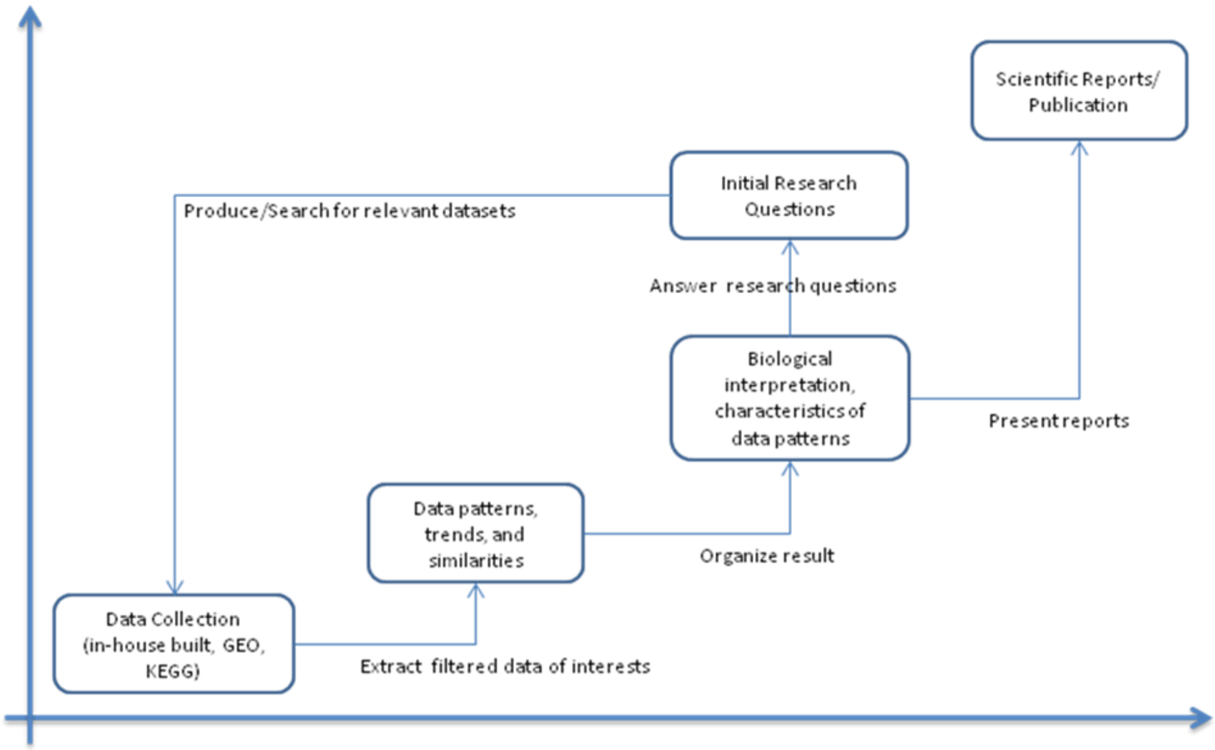

Figure 3. Information Processing in Use Case 1 - Biomedical Research

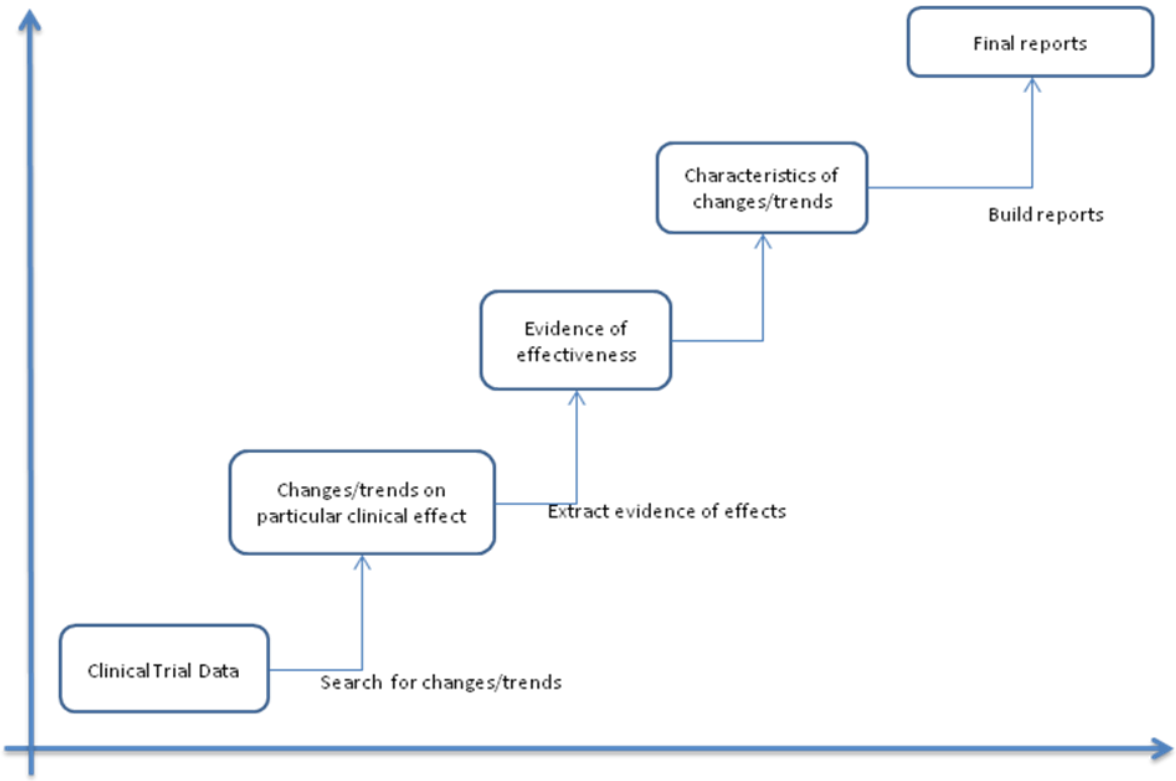

Figure 4. Information Processing in Use Case 2 - Clinical Trial

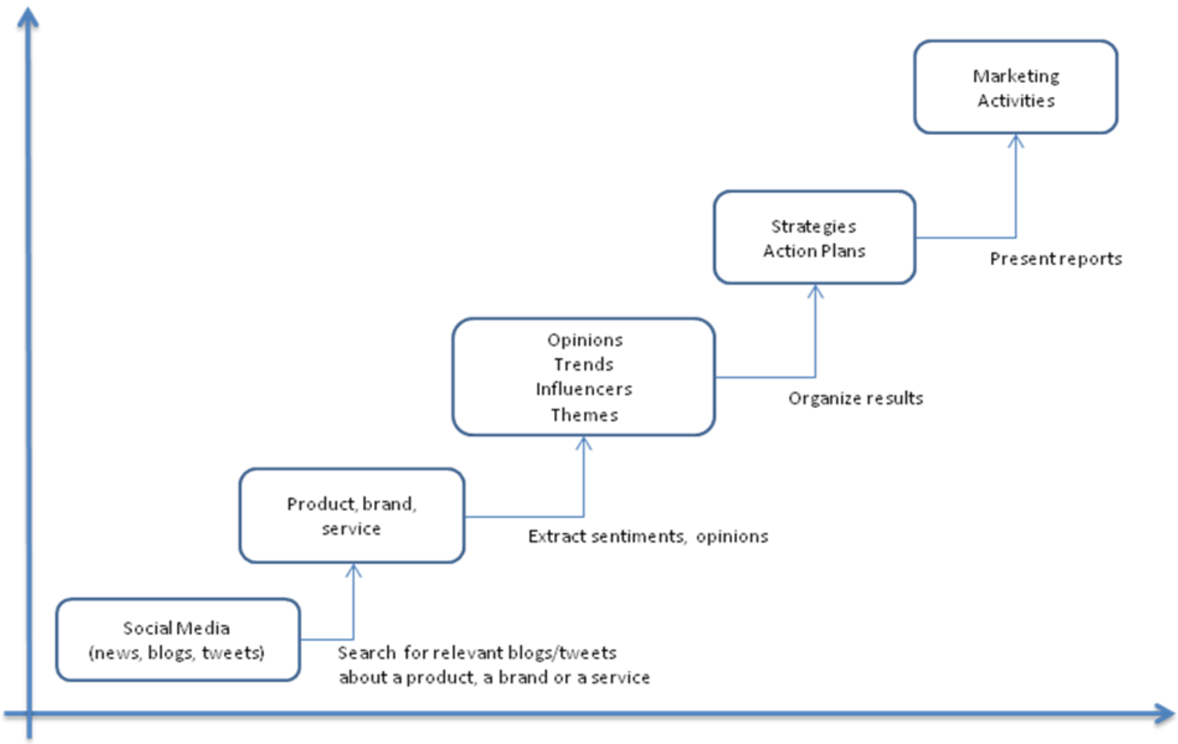

Figure 5. Information Processing in Use Case 3 - Social Media Monitoring

3.2 Collaborative Sensemaking Framework

Sensemaking extends far beyond individuals making sense of their information spaces. Three use cases of Dicode have all claimed that collaboration is highly desired (see D6.2.1, D6.3.1 and D6.4.1). A group of people need to work together to understand complex issues, combining information from multiple data sources and bringing together different experience and expertise towards a shared understanding.

However, there has been little exploration of how sensemaking takes place in collaborative work. Studies reported related to sensemaking are from different domains, perspectives or focuses. Ntuen [5] has studied collaborative sensemaking in military coalition operations, where a group of people with different worldviews are collectively engaging in making sense of chaotic and ambiguous situation. Lee and Abrams [4] have further explored sensemaking regarding to collaboration, which could entail innovation on at least two levels: joint learning about how to collaborate and coordinate work, and joint learning in how to represent and instantiate a design that does not yet exist. Qu and Hansen [10] have proposed a conceptual model of collaborative sensemaking, which distinguishes between shared representation and shared understanding. They also argued that collaborators can develop a shared understanding by examining, manipulating and negotiating external representations. Paul and Reddy [6] have discussed a framework of collaborative sensemaking during CIS (Collaborative Information Seeking) activities and design implications for supporting sensemaking in collaborative information retrieval tools.

From the above literature, we can see that collaborative sensemaking covers a wide range of collaborative processes, which is far more than the aggregation of individual sensemaking. There is no one universal framework to describe this complex process. We find Paul and Reddy's framework (Figure 6) more relevant to our studies because it links individual sensemaking and collaborative sensemaking, and defines triggers and characteristics of sensemaking.

This framework highlights some important factors that **trigger** collaborative sensemaking during a CIS activity, namely, ambiguity of information, role-based distribution of information, and lack of expertise. This framework shows that CIS activities are often initially split into tasks/sub-tasks and sub-tasks are performed by different group members, depending on their roles and expertise. Roles can be organizational or might be assigned informally. During individual sensemaking, action awareness information is shared, i.e., group members keep each other aware of what they are doing. The framework highlights that CIS activities often involve individual information seeking and sensemaking and then lead to collaboration.

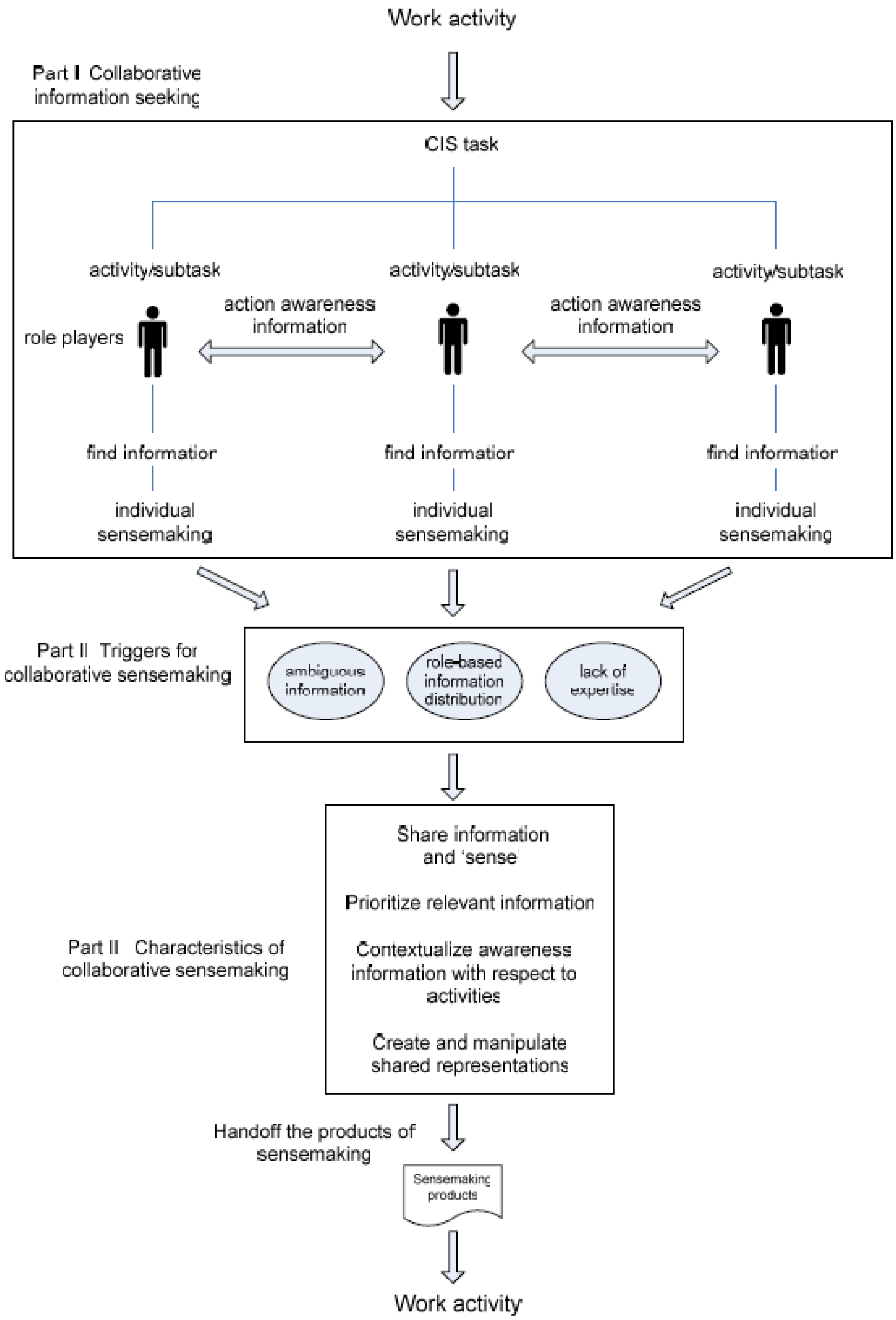

Figure 6. Collaborative Sensemaking Framework (Paul and Reddy, 2010)

The framework lists some **characteristics** of collaborative sensemaking, namely, prioritising relevant information, sensemaking trajectories, and activity awareness. Prioritising the 'right' pieces of information as relevant enhances group sensemaking. Knowing the "path" that a group member followed to make sense of information helps other group members' sensemaking. Such paths are called *sensemaking trajectories*. Group members share and make sense of information, they create shared representations to store the information found and the sense made of that information. The characteristics and the triggers of collaborative sensemaking identified in this framework provide us a guideline to understand the demand of collaboration in Dicode use cases [Table II].

	Description in the model	Examples in Dicode use cases	
Triggers	Ambiguous information; Role-based information distribution; Lack of Expertise	UC1	Acquire expert support (e.g., a researcher needs the support of other researcher on whether his/her interpretation of the result is significant)
		UC2	Clarify results with other experts (e.g., a clinician wants to know whether his/her result is aligning with the result of other trial sites.)
		UC3	Transfer knowledge (e.g., social media analysts)
Characteristics	Prioritizing relevant information; Sensemaking trajectories; Activity awareness	UC1	Get opinions from other scientists about choosing right datasets, databases, tools
		UC2	Be aware of results from other trial sites; Collaboratively diagnose complex cases
		UC3	Be aware of activities of other parties; Collaboratively transform data results to valuable insights.

Table II. Collaborative Sensemaking in Dicode Use Cases

3.3 Generic Conceptual Architecture for Dicode

Dicode's generic conceptual architecture is reflecting Dicode's solution to support the characteristics (both differences and commonalities) of big data analytics derived from the Dicode use cases and sensemaking frameworks. As shown in the architecture diagram (Figure 7), the Dicode Conceptual Architecture is supporting bottom-up sensemaking process (from Input data to Knowledge Product) through three types of services:

- **Data-centric services**
  Data-centric services exploit large data processing technology to meaningfully search, analyze and aggregate data from heterogeneous data sources. The input of the data-centric services is structured and/or unstructured data from heterogeneous data sources. The output of data-centric services is searched or filtered information, discovered patterns or lists etc. The data-centric services aim to improve the processes of individual sensemaking.

- **Collaboration-centric services**
  Collaboration-centric services support people and their interaction by capturing and sharing resources, opinions, arguments and comments among participants, so to facilitate the collective understanding of the issues related to data analysis. The input of the collaboration-centric services could be the output of data-centric services as well as the interactions (comments, arguments and discussions etc.) among all parties. The knowledge

product (hypothesis, strategies etc.) should be the outcome of their interaction. The collaboration-centric services aim to support collaborative sensemaking.

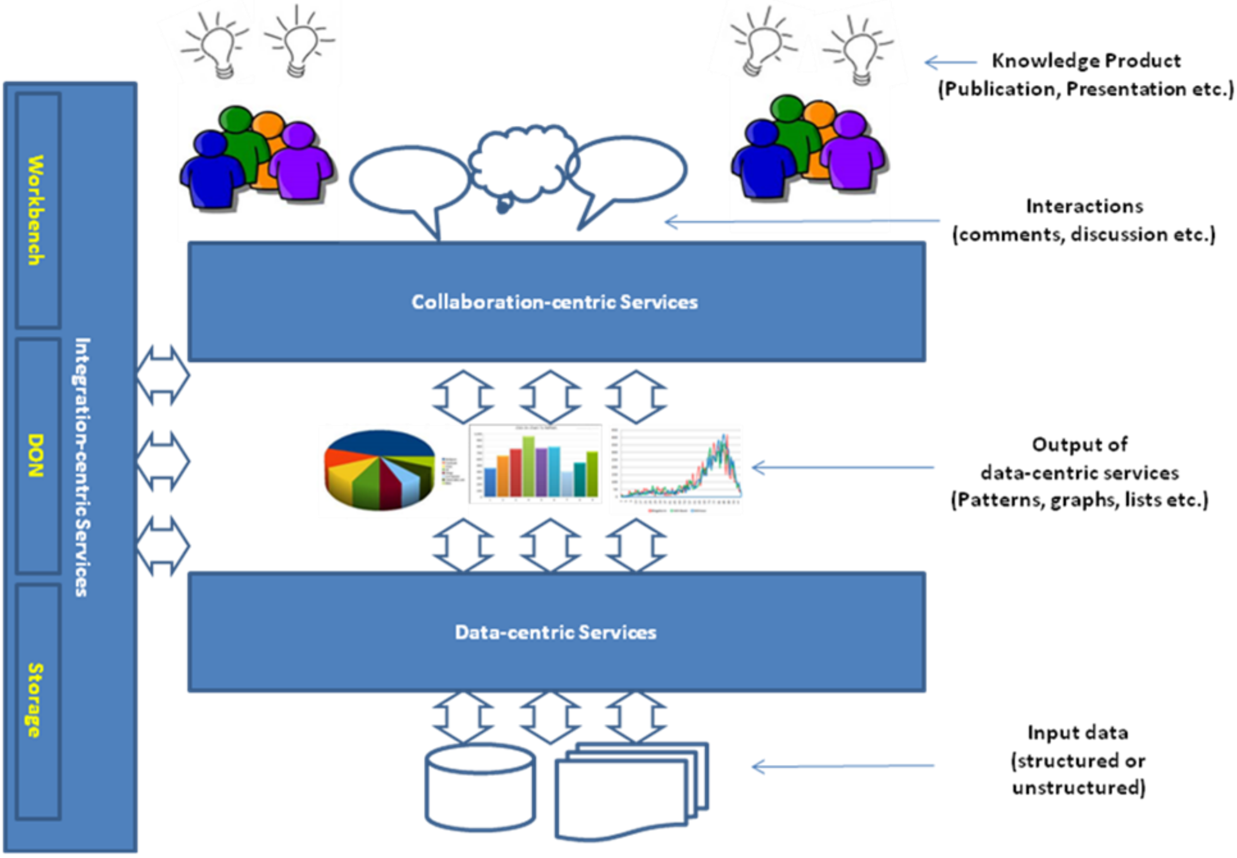

Figure 7. The Conceptual Architecture for Dicode

- **Integration-centric services**
  Integration-centric services support data-centric services and collaboration centric-services. Their aim is to ensure and facilitate the seamless – from both the conceptual and the technical point of view – interoperability and integration of the independent services developed. Related functions include user registration, service registration and data storage so that users can access services in a holistic way through underlying annotation and augmentation of the services. The integration-centric services, which include Dicode workbench, Dicode ONtology (DON) and Storage service, remain the same for all use cases.

  - *Dicode workbench*
    The Dicode Workbench provides functions of user management and service management. It allows users to assemble their Dicode Workspaces from a unified web interface. Through the Dicode Workbench, users can access different services (data-centric services and collaboration-centric services) developed within the Dicode project via widgets.

  - *Dicode Ontology (DON)*
    DON is a multi-layered ontology, designed to address requirements from multiple use cases that involve sensemaking. DON is used as a common vocabulary among services

and service developers for enhancing the functionality of Dicode services. DON plays a crucial role to facilitate the integration and interoperability of services. The main idea is that some features of the services will be annotated using concepts included in the DON. The information about services and their annotations will be maintained in a central registry (Dicode Service Registry - DSR). This registry will be available for the rest of the components of the Dicode environment through a REST interface.

- *Storage service*
  Storage Service is to provide Dicode users with a permanent and reliable storage place to keep files accessible. The service will be as generic as possible to allow storing any kind of files (text plain, doc, pdf, html, xml, json, zip, etc.). The service provides mechanisms to upload files and retrieve them by using RESTful services. Additionally, meta-data information about files will be also stored to facilitate their search and location by search engines or services. These meta-data will contain information such as type of file (pdf, html, xml, etc.) or type of content (Dynamika report, DNA sequence, etc.).

Although the Dicode solution is supporting bottom-up process sensemaking driven by data, a group of users may start using the collaboration-centric services first (e.g. to discuss about available or interesting data sources, or even to specify important parameters to be taken as input by the data-centric services), and then engage the data-centric services. Generally, there is a loop between these two service categories, which is reflected by the bi-directional arrow between these two types of services.

A detailed list of services developed in the context of Dicode can be found at Dicode Wiki[2]. The list presents the name of the service, the use cases it is intended to be used, along with a description, which indicates what it does, under which circumstances to use it and provides a characteristic example scenario.

4 Use Case 1: Clinico-Genomic Research Assimilator (CGRA)

The update of this use case has been provided in D6.2.1, which highlighted that biomedical research has become increasingly interdisciplinary and collaborative in nature. The vast amount of the data available and the ever increasing specialised resources show that the way forward is to form biomedical research collaboration teams to address complex research questions [2]. Such interdisciplinary teams would better meet challenges relative to various problems such as how to store, access, analyse and integrate multiple types of data or, how to work with multiple databases simultaneously; or even, how to make data accessible and usable to life sciences researchers.

4.1 Instantiated Conceptual Architecture of Use Case 1

To support CGRA use case, Dicode is to provide an integrated tool for a multidisciplinary biomedical community with members ranging from biologists to bio-informaticians, as well as data-centric and collaboration-centric services to assist individual members to deal with increasing volume and diversity of data sources. In the instantiated architecture shown in Figure 8, the data sources are specific in biomedical research so the services are chosen to deal with data processing and analysis in this field, such as R service, Subgroup Discover and GEO

[2] https://wiki.dicode-project.eu/display/DIC/Evaluation+of+Dicode+Services

Recommender. Workspace, Logbook and Forum summarization services are supporting the collaboration and interaction among the users.

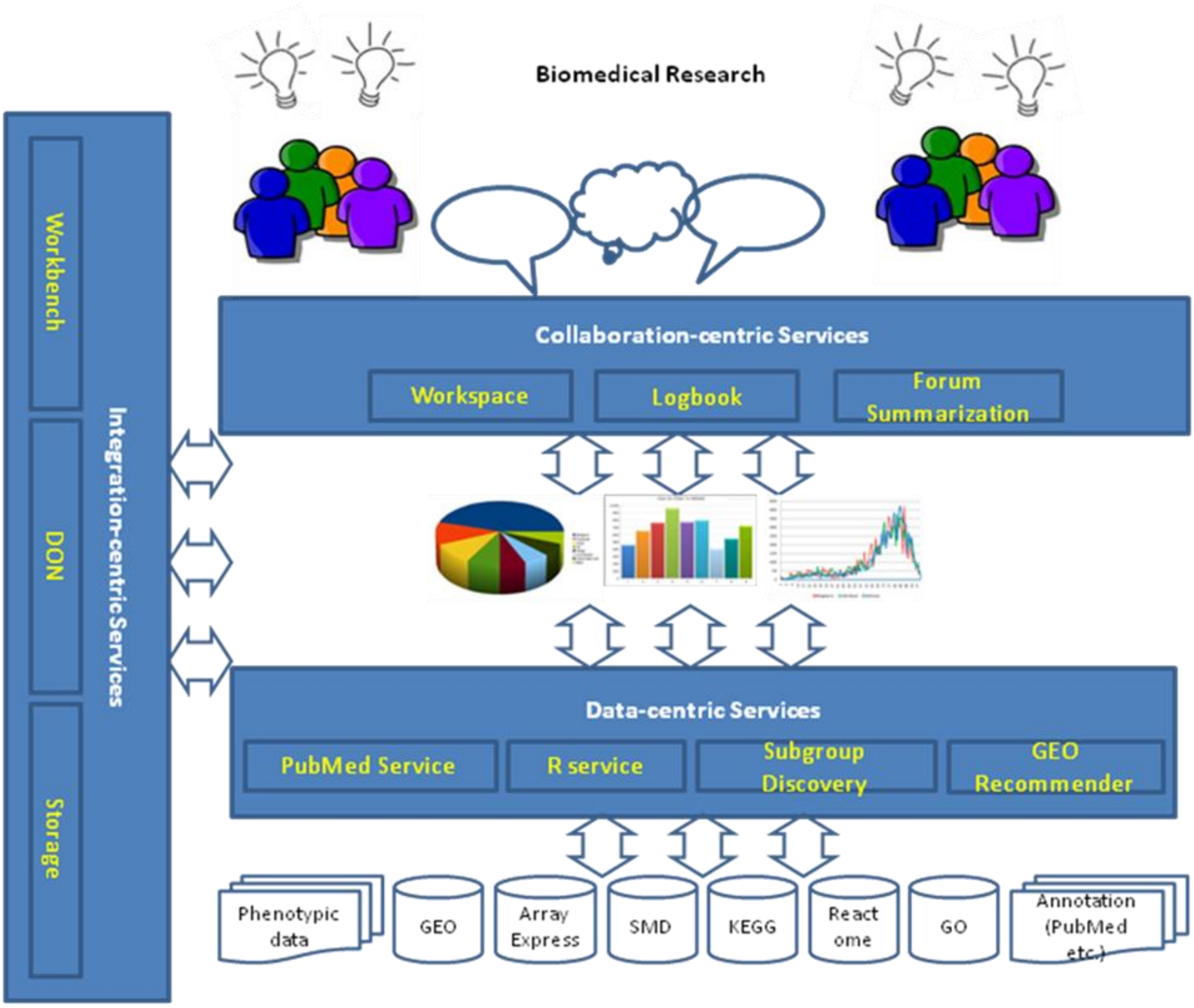

Figure 8. Dicode Conceptual Architecture - Use Case 1

4.2 Dicode Services and Usage Scenario of Use Case 1

Here we present an example on how Bioinformatics researchers benefit from Dicode platform for their work.

Sarah (PhD student), James (Postdoctoral Researcher) and John (Professor, supervisor of Sarah and James) are three researchers from a Breast Cancer research institution. They have conducted some studies on a small sample-size gene-expression microarray breast cancer dataset. The analysed result is not satisfactory but they believe that some extra datasets from public resources, such as GEO (Gene Expression Omnibus) with the same pathology characteristics can augment their sample size and allow them to identify some extra statistically significant genes.

All are using ***Dicode workbench*** *to coordinate their work and support their research. Each of them has an account on the Dicode workbench and this enables them securely share their work. Using the* ***Storage service,*** *both Sarah and James have uploaded some graphs and data on what they have found out from their studies.*

*Working towards a publication, Sarah has added the **PubMed service** to their Dicode workbench. Using this service, she discovers relevant publications which address similar biological questions and may be used to justify their sample size choice. The result from PubMed tool has been recorded and can be seen by James and John at any time.*

*When making their decision on sample size, they want to search in appropriate forums for extra evidence/ arguments. Using **Forum summarization service**, James quickly checked a forum (such as "Cross-validated", a statistical analysis forum) and see if the topics discussed are relevant to their research.*

*Having a brief idea about their sample size, the team "meet" in the **Collaboration workspace** to brainstorm their ideas and their opinions (agree, disagree, comments, ideas, support documents etc).*

*To understand more about James' work, Sarah asks James to upload his R-script as she wants to know whether a few arguments (lines of code) could be rearranged. Using the **R service**, James run his R-script with some new arguments and a new graph is easily produced for everybody to assess the new strategy and decide on the significance of the results.*

*After a collaboration session, James has collected enough information about the data and sample size he needs for his task. James then launches the **GEO recommender** service to get the datasets. He types in the request describing the data and also the methodology he will apply. All qualified datasets are provided in a list.*

*From the list of recommended datasets, Sarah wants to find the functional interpretation of expressed genes in two datasets and compare them. She first launches the **R service** to identify expressed genes. In the second step, she uses **Subgroup Discovery** service, which provides a list of subgroups describing the expressed according to their molecular function and their role in biological process, which has shown a good match to their previous findings.*

***Dicode Logbook** has recorded their current status and evolution of past and on-going research activities. They all can look into a particular discussion they have gone through.*

5 Use case 2: Trial of Clinical Treatment Effects (TCTE)

The update of this use case has been provided in D6.3.1 to support broader clinical trials, not just for Rheumatoid Arthritis. As the business domain has been expanded, the goal of this use case is now set 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. The collaboration in a clinical trial, especially in multi-site clinical trial is not in practice but highly needed because:

- it can improve the quality of diagnoses by enabling multiple experts examine one single case or comparing similar cases;
- it will speed up the process and save the cost of trial by sharing the result simultaneously and monitoring the overall process as a whole.

5.1 Instantiated Conceptual Architecture of Use Case 2

In the instantiated architecture appearing in Figure 9, the data sources are specific in clinical trials, i.e. the DICOM files and Dynamika Reports. Consequently, the services chosen to deal with data processing and analysis in this field are Subgroup Discovery and Augmentor (together with its background services, namely Semantic Augmentation, Semantic Query and Entity Summarization).

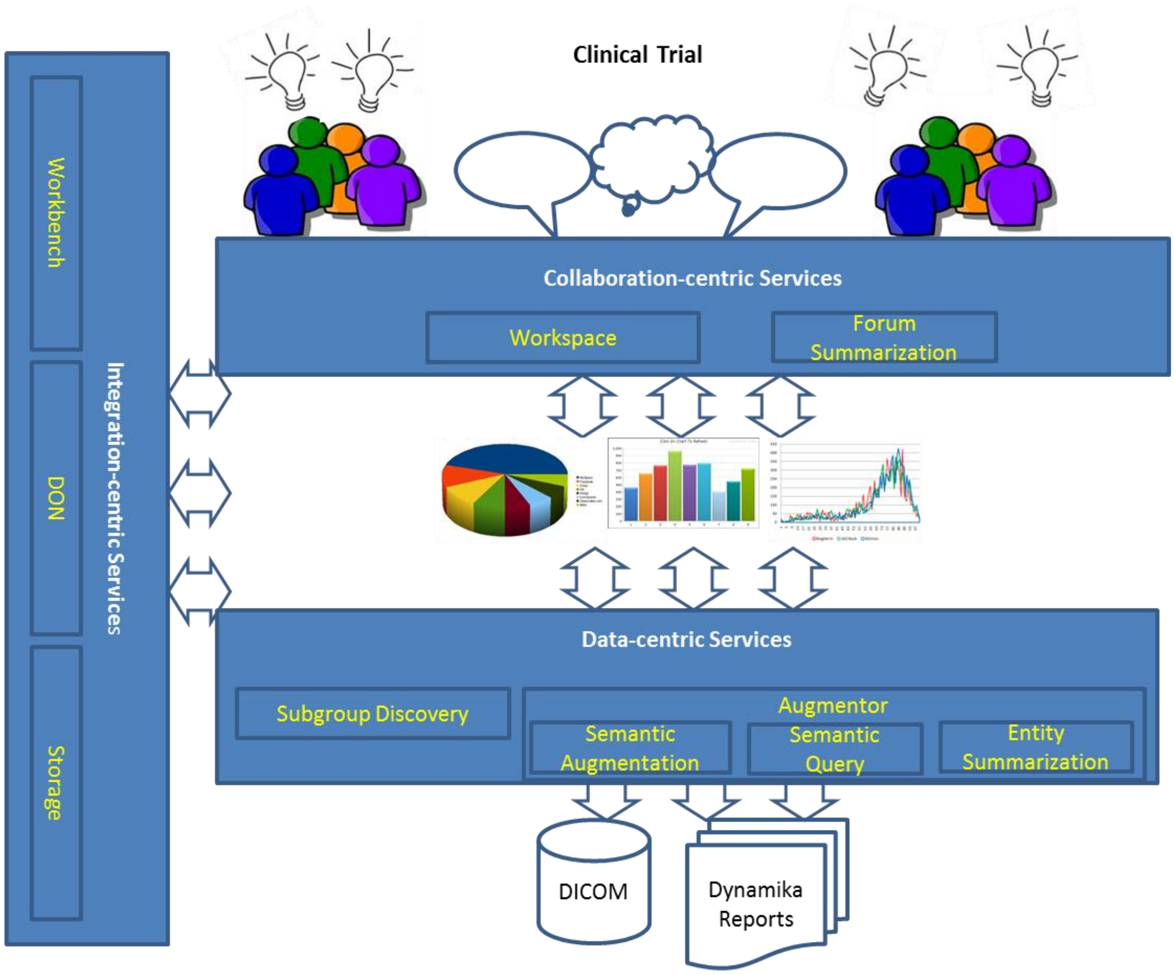

Figure 9. Dicode Conceptual Architecture - Use Case 2

5.2 Dicode Services and Usage Scenario of Use Case 2

Here we present an example scenario on how a Clinical Trial Team benefit from Dicode platform for their work.

> *A group of radiologists and clinicians are sharing their findings from of a drug trial conducted in three sites worldwide: US, UK and Australia.*
>
> *The **Dicode workbench** allows all parties to collaborate during the whole process. All of them have an account with the Dicode workbench, which enables them to securely share their data and their discussions.*
>
> *For this drug trial, all clinicians and radiologists are working on one **Collaboration workspace**. The leading clinicians in these three sites have uploaded their current progress together with some supporting documents using the Dicode **Storage Service**. Authorised clinicians and radiologist can see the shared data.*

The leading clinician, Chris from UK, has not seen any good results since the trial started. However, after reading reports from the US, he found some cases showing significantly promising results. He asked the leading clinician in US, Ann, to share some Dynamika reports related to those cases.

Ann and Chris look at the shared reports over Dicode platform, discuss their opinions in the ***Collaboration workspace*** *on why these cases are showing promising results. They also notice from* ***Augmentor*** *Service, that the comments of Dynamika reports have shown that a particular feature of the patient has been frequently mentioned in the comments. The Augmentor lists the reports with similar comments for Ann and Chris to conduct further investigation.*

While examining a particular case in Australia, the leading clinician, David, notices the discussion happening between Ann and Chris in the collaborative workspace. He joins the discussion with some data he collected from Australia. Although there are only a couple of cases showing promising results like the data shown from US cases, David believes that it is possible to find similar samples to produce similar result.

He launches the ***Recommender service*** *to find such samples. The service returns a set of reports from US, which share similar content in Dynamika reports. The discovered reports have provided valuable information for all of them. Based on the discovered reports, both Chris and David have redesigned their trial plan and produced a similar result to US. This has shortened their trial process about 12 months.*

6 Use case 3 - Opinion Mining from Unstructured Web 2.0 Data (OPUWD)

The update of this use case has been provided in D6.4.1 to address the importance of collaboration beyond professionals within a Public Relationship (PR) company for social media monitoring and engaging process. In a fast-changing world, where social media is influencing consumer demands, a successful media engagement strategy depends on the collaboration of all relevant parties – public relations, brand, media and marketing [3].

A collaborative environment is expected to facilitate sensemaking and decision making. Through such solutions, all parties will be much more effective in locating, retrieving and meaningfully interact with relevant information.

6.1 Instantiated Conceptual Architecture of Use Case 3

In the instantiated architecture shown in Figure 10, the data sources are specific in social media monitoring: dedicated news feeds, tweets and blogs. Consequently, the services are chosen to deal with data processing and analysis in this field, such as topic, influencer and sentiment analysis, etc.

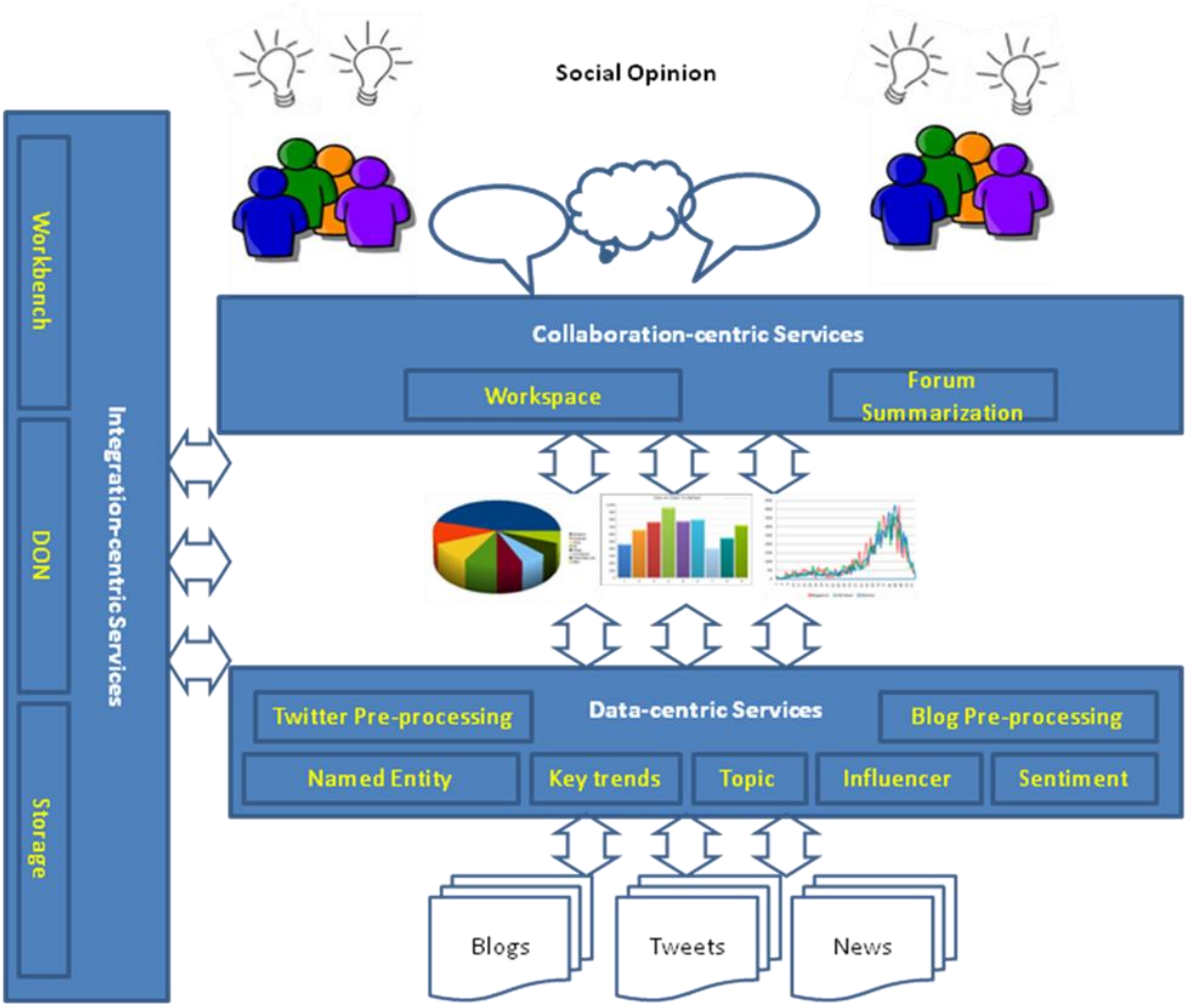

Figure 10. Dicode Conceptual Architecture - Use Case 3

6.2 Dicode Services and Usage Scenario of Use Case 3

Here we present an example scenario on how a company benefit from Dicode platform on their new product launching process.

> *A car manufacturer is launching a new product. In this process, three main parties are involved. One is a Brand Manager (Frank) from the marketing department of the company. The second one is a Social Media Analyst (Alice) working in a marketing consultancy. The third one is Social Media Engager (Natalie) working in a public relations agency responsible for social media engagement.*
>
> *The* ***Dicode workbench*** *allows all three parties to collaborate during the whole process. Frank has a question about first consumer experiences with the new product in the social web and gives a briefing to Alice.*
>
> *Alice starts analysing the web and updates the results in the* ***Collaboration workspace****. She watches over social media and provides advice to the Brand Manager. She detects the significant conversations and news articles with the* ***Topics service*** *and looks for insights as a basis for product development or communications from the blogs and tweets. If she wants to get deeper information on relevant tweets detected, she can use* ***Key trends service*** *to show trends on*

Twitter, such as the top links for a certain day posted by twitter users. She can also use Phrase extraction service with the pre-trained sentiment model to monitor positive or negative sentiments that are expressed in connection with the brand.

Frank recently has come across an e-mobility forum and has identified this forum as an interesting source for the car manufactory's business. Frank checks this forum using ***Forum Summarization*** *Service and identifies the most relevant topics. Based on his findings, he briefs Natalie to engage in this forum.*

Frank can directly ask questions and / or give advice to control the research conducted by Alice. Natalie can access the results that Frank and Alice have provided to understand more about the current opinions from social media.

In parallel, Frank can start thinking about marketing activities to promote the product or to change packaging and/ or communications. He can pre-align the activities with further involved parties in and out of the company. At the meantime, Frank can quickly brief Natalie on engaging with identified blogs.

7 Conclusion

In this deliverable, we have presented the updated Dicode approach including the methodology, the conceptual architecture and usage scenarios, which show how the three use cases can benefit from the developed Dicode workbench. This update is made by acquisition of new knowledge and by reflection on the achievements to date: the development of Dicode services and user trial results as described in D6.2.1, D6.3.1 and D6.4.1.

The Dicode methodology has brought use case partners and technical partners together and enabled the development of Dicode services in an incremental manner. This methodology reinforces the equal partnership between end users and technical partners. This mutual understanding is crucial in the exploration of an innovative solution for fostering synergy between human and machine reasoning.

Underpinned by sensemaking models, the commonalities and differences of three use cases have been analysed to inform the generic conceptual architecture for mastering data-intensive collaboration and decision making. The Dicode conceptual architecture provides a common framework to guide the direction of investigation and experimentation in the project. The instantiation of the architecture for each use case has deepened our understanding in the likely impact of unique characteristics of the associated data and processes. It has also shown that the Dicode integrative approach can be applied to different domains and can be evolved with increasing understanding and contributions towards big data analytics.

References

[1] Fisher, D et al. (2012). Interactions with Big Data Analytics, Interactions Volume XIX.3, May+June 2012, ACM.

[2] Karacapilidis, N., Tzagarakis, M., Christodoulou N. and Tsiliki, G. (2012). Facilitating and Augmenting Collaboration in the Biomedical Domain. International Journal of Systems Biology and Biomedical Technologies, Vol. 1, No 1, 2012, pp. 52-65.

[3] Karacapilidis, N., Loeffler, R., Maassen, D. and Tzagarakis, M. (2012). Augmenting Social Media Monitoring through Human Collaboration. In: Proceedings of the 16th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2012), San Sebastian, Spain, September 10-12, 2012 (to appear).

[4] Lee, C.P., Abrams, S. (2008). Group Sensemaking. CHI 2008 Sensemaking Workshop. Florence, Italy, 2008.

[5] Ntuen, C. A., Munya, P. and Trevino, M. (2006). An Approach to Collaborative Sensemaking Process. Proceedings of the International Command and Control Technology Symposium (ICCRTS) (Cambridge, UK, 2006). Command and Control Research Program (CCRP).

[6] Paul, S.A. and Reddy, M. (2010). A Framework for Sensemaking in Collaborative Information Seeking. Proceedings of 2nd International Workshop on Collaborative Information Seeking at CSCW 2010, Savannah, GA.

[7] Pirolli, P. and Card, S.(1999). Information Foraging. Psychological Review, 106, pp. 643-675.

[8] Pirolli, P. and Card, S. (2005). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. Proceedings of International Conference on Intelligence Analysis.

[9] Russell, D. M., Stefik, M. J., Pirolli, P., & Card, S. K. (1993). The cost structure of sensemaking. Paper presented at the INTERCHI '93 Conference on Human Factors in Computing Systems, Amsterdam.

[10] Qu, Y. and Hansen, D. L. (2008) Building Shared Understanding in Collaborative Sensemaking. Proceedings of the Workshop on Sensemaking at the Conference on Human Factors in Computing Systems (CHI), ACM.

[11] Yang-Turner, F. and Lau, L. (2011). A pragmatic strategy for creative requirements elicitation: From current work practice to future work practice, Requirements Engineering for Systems, Services and Systems-of-Systems (RESS), pp.28-31, ACM.