FP7-ICT-2010-270253

Integrate

Driving excellence in Integrative Cancer Research through Innovative Biomedical Infrastructures

STREP
Contract Nr: 270253

6-Monthly Progress report # 1

Feb - July 2011

Due date: August 2011
Actual submission date: September 2011

Beneficiary name	Beneficiary short name	Country	Date enter project	Date exit project
Philips Research	Philips	NL	M1	M36
Breast International Group	BIG	B	M1 M1	M36 M36 M36
Foundation for Research and Technology – Hellas	FORTH	GR	M1	M36
Custodix	Custodix	B	M1	M36
Institut Jules Bordet	IJB	B	M1	M36
Universidad Politechnica de Madrid	UPM	SP	M1	M36

Start date of Project: 01 February 2011 Duration: 36 months

0 Document info

0.1 Author

Author	Company	E-mail
Anca Bucur	Philips	anca.bucur@philips.com
Kristof De Schepper	Custodix	kristof.deschepper@custodix.com
Alexandre Irrthum	BIG	alexandre.irrthum@bordet.be
David Pérez del Rey	UPM	dperez@infomed.dia.fi.upm.es
Manolis Tsiknakis	FORTH	tsiknaki@ics.forth.gr
Konstantinos Marias	FORTH	kmarias@ics.forth.gr
Debora Fumagalli	IJB	debora.fumagalli@bordet.be

0.2 Document data

Keywords	Interim activity report
Editor Address data	Name: Anca Bucur Partner: Philips Address: HTC 34 (3-058) 5656 AE Eindhoven The Netherlands Phone: +31 40 27 Fax: +31 40 27 E-mail: anca.bucur@philips.com
Delivery date	September 2011

0.3 Distribution list

Date	Issue	E-mailer
29/09/2011	1.0	Loukianos.GATZOULIS@ec.europa.eu
29/09/2011	1.0	fp7-integrate@listas.fi.upm.es

Table of Contents

1 Workpackage progress of the period

1.1 WP1 User needs and requirement (Lead: IJB)

1.1.1 Objectives (of the reporting period)

The main objectives of the WP for this period are to identify the users and their needs, to define and prioritize comprehensive user scenarios on which use cases will be based, and to define legal and regulatory requirements.

1.1.2 Status/progress towards objectives WP1 (per Task)

Task 1.1 Identification of the users and their needs

User needs for the INTEGRATE environment were initially gathered through interviews and discussions with leading oncologists and researchers from the NeoBIG research program that promotes data sharing in the contect of neoadjuvant breast cancer therapy.

More detailed user requirements for the INTEGRATE environment were then elicited from a larger panel of potential end-users and advisors from BIG and IJB, including oncologists, translational researchers, clinical trial administrators, legal advisors, health IT specialists and data analysts.

Working reunions were held weekly within BIG and Institut Jules Bordet, and regularly through teleconferences and face-to-face meetings between all members of the consortium. The opinions of external advisers were also solicited and gathered during some of these meetings and during other events such as conferences.

The different categories or roles of end users of the INTEGRATE platform have been identified during this reporting period. These roles include clinicians, core laboratory staff, administrators and researchers from academia and pharmaceutical companies. Sub-categories of users have also been identified. For example, clinician has been sub-divided into investigator, radiologist, pathologist, clinical research nurse, etc.

Access requirements and access rights associated with these user roles have been thoroughly discussed within IJB/BIG and a document describing these requirements and rights has been drafted.

The main product of the activities of IJB toward completion of this task is deliverable D1.1 “User needs and specifications for the INTEGRATE environment”, which has been submitted to the European Commission.

Task 1.2 Definition of user scenarios

A large part of the effort during this reporting period has involved definition of the user scenarios.

Scenarios have been defined for the following functions of the platform:

- Molecular screening

- Biotracking (biospecimen tracking)
- Retrospective use of clinical, molecular and imaging data, including predictive model building
- Central review of pathology images
- Pilot for interaction with the electronic health records

Documents describing these scenarios have been drafted and distributed within the consortium.

An important sub-task here to ensure that user scenarios can be exploited correctly is agreement on a shared, unequivocal vocabulary. To this end, a glossary of terms from the user scenarios has been constructed and placed on the project wiki.

Task 1.3 Legal and regulatory compliance requirements

Legal and regulatory compliance requirements were given considerable consideration throughout the reporting period. Legal counsellors from Institut Jules Bordet and BIG, as well as clinical trial specialists, participated to the working reunions related to user requirements and user scenarios, which allowed early identification of potential legal and regulatory issues.

Several meetings specifically devoted to the discussion of these matters were held within IJB and BIG. External advisors with a relevant experience in data sharing for clinico-genomic trials (Sage bionetworks, I-SPY…) were also identified and contacted and their recommendations were gathered during face-to-face and teleconference meetings.

An initial draft of the deliverable D1.3 “INTEGRATE legal, ethical and regulatory requirements” has been completed. This document presents all the relevant topics that have been identified in this respect, including topics related to data protection, informed consent, intellectual property rights and contractual matters.

1.1.3 Deviations from the DOW and corrective actions

There are no major deviations from the DOW.

1.1.4 Planning next period

Work for the upcoming period will focus on refining of the user scenarios. To this end, documents describing relevant procedures from previous NeoBIG clinical trials have been identified and this information will be incorporated in the existing scenarios. Lessons learned from ongoing pilot studies (see below) will also help to refine the scenarios.

Legal and regulatory compliance requirements will also receive considerable attention during the next period.

1.2 WP2 Architecture and integration (Lead: Custodix)

1.2.1 Objectives (of the reporting period)

Start defining an initial architecture which integrates the different components and tools (modules and services) provided by the INTEGRATE project. This contains the initial definition of use cases, security/component/data/information/... models and semantic solutions, based on the provided stakeholders scenarios and requirements defined in WP1. Also an Identification and evaluation needs to be made on the relevant standards and technologies for the INTEGRATE state-of-the-art document.

1.2.2 Status/progress towards objectives WP2

Task 2.1 Identification and evaluation of relevant standards

- A first state-of-the-art draft is generated for deliverable D2.1 (month 9) containing research in following topics:
  - Relevant ontologies and vocabularies
  - Semantic repositories
  - Automated reasoning in the semantic web
  - Ontology mediation, alignment and merging
  - Ontologies for the life sciences
  - Data and ontology sources
  - Query languages for semi-structured data
  - Security and privacy standards

Task 2.2 Inventory of re-useable/available relevant solutions and components

As functional components are identified, the first re-useable/available relevant solutions and components are identified. However this work has only been starting at the very end of the reporting period.

Task 2.3 Design and implementation of the INTEGRATE reference architecture

- Several brainstorm meetings were held, defining the core influential aspects of the INTEGRATE architecture (i.e. semantic approach, information models, possible interface technology). This work will continue during the next reporting period.
- A first draft of a technical use case document was created based on the scenarios and requirements of WP1. Developing the technical use cases is part of the functional decomposition process.
- A first draft of the INTEGRAGE component model based on the use cases was created.

Task 2.4 Security for dynamic collaborative environments

Work has been started on translating security requirements into a model for the security solution.

1.2.3 Deviations from the DOW and corrective actions

N/A

1.2.4 Planning next period

- Design of the initial architecture.
- Reviewing and extending the first draft of the use cases.

- Updating the component model using the new version of the use cases.
- Building the security model based on the user scenarios and requirements.
- Completing and reviewing state-of-the art deliverable.
- Re-usable/available relevant solutions and components.

1.3 WP3 – Data Models and Interoperability (Lead: UPM)

1.3.1 Objectives (of the reporting period)

The main objective of this WP is to enable interoperability within the INTEGRATE environment, providing services such as data extraction, transformation and mapping among data models of clinical infrastructures. For the reporting period, the objectives were mainly focused on the core dataset (Task 3.1) and information models (also called common data models) (Task 3.2).

The core dataset is the shared vocabulary, including the corresponding relationships (also known as ontologies), required to accomplish the semantic interoperability among heterogeneous systems. To integrate heterogeneous data models from different sources, mappings linking the core dataset and common data model concepts are required. The semantic interoperability layer will use these mappings to provide a uniform and semantically interoperable platform for Electronic Health Records and Clinical Trial data (including external sources).

1.3.2 Status/progress towards objectives WP3 (per Task)

Task 3.1 Definition of the semantic core dataset

Within the core dataset task, a set of relevant domain concepts, describing the semantics of the clinical domain, should be identified. Since INTEGRATE aims to reuse available terminologies, different standardized terminologies have been analyzed as core dataset candidates, i.e. SNOMED CT, ICD-10, MedDra, LOINC and MeSH. The analysis suggests that our platform will require various terminologies to cover the scenarios described by the user requirements. Although SNOMED CT may provide the majority of the core dataset concepts, areas such as adverse events or laboratory test results will require the use of MedDra and LOINC terminologies. The overlapping among these terminologies will be handled by selecting a “default” vocabulary for each area.

To provide an environment allowing reasoning required to perform complex queries over a common vocabulary, different ontology languages, such as RDF, SKOS or OWL, have been analyzed to store the core dataset. Existing projects claim to provide tools to automatically transform certain terminologies to such languages. Database models have been also analyzed to store the core dataset, offering an improved performance but more complex or limited reasoning.

A preliminary set of terms to develop the core dataset have been identified manually by the users and has being compared with available terminologies. Results and comparison with other core dataset from similar projects suggest that we should minimize the use of post-coordination to describe new concepts. Automatic methods to extract core dataset terms from case report forms and eligibility criteria have been also tested with promising results.

Task 3.2 Definition of the information models of the clinical and research infrastructures

Common data models are required for mappings that will be developed in task T3.3, bridging core data set concepts with concepts of such models. They aim to provide a canonical view, reflecting the content and the structure of each data source. The following features, among others, are being taken into account to design the common data model of INTEGRATE: modeling capabilities to store data from requirements, performance, potential for reasoning and minimizing structural modifications when new sources are integrated. There is a trade-off between the last feature, requiring a simple design with dynamic updates for new sources, and performance. Too simplistic solutions will have lower performance, while complex schemas will require more changes in the future. A set of data models candidates is being currently analysed: i2b2 (informatics for integrating biology & the bedside), OMOP (Observational Medical Outcomes Partnership) and HL7 RIM (Health Level 7 Reference Information Model).

For the first implementation of the INTEGRATE platform, tools to Extract Transform and Load (ETL) original data sources into the common data model will be required. Although schema heterogeneities could be solved by pure federated approaches, data transformations require such tools to avoid problems mainly with performance and reliability. A state of the art has been carried out, regarding ETL tools available. Pentaho Kettle and Talend are open source projects that can be used for this task within INTEGRATE. Both have a large community of users and have been previously used within biomedical integration projects.

Task 3.3 Semantic formalism, mapping tools and mapping implementations

This task aims to identify the requirements to link core data set concepts, EHR sources and the clinical trial management system. Such sources will be stored at each node following the common data model from the previous task. Mapping requirements are being considered to design the core data set and common data model. We are also analyzing user scenarios and the corresponding use cases to identify requirements for reasoning that such mappings should meet.

The mapping implementation of the INTEGRATE platform will need the input of domain experts to link concepts and data models. Schema transformations, driven for such mappings, will be performed by the semantic interoperability layer to provide a uniform view of the data sources within INTEGRATE.

Task 3.4 Design and implementation of the semantic interoperability layer

The semantic interoperability layer will execute the mappings during the data extraction phase, instantiating the semantic concepts with patient data and/or clinical trial data, enabling the linkage between the patient data in the INTEGRATE repositories and the patient data in the existing clinical and research systems. This capability will allow tools and services of INTEGRATE to access all the necessary data out of the INTEGRATE repositories and of the relevant existing research and clinical systems in a semantically-aware and uniform way.

The core dataset, the common data model and the mapping approach will be included within the semantic interoperability layer. During the reporting period, the suitability of the core dataset and common data models to perform the required integration have been the focus of this task. In addition we have carried out an analysis of current query languages that can be used to retrieve data from the platform: RDQL, RQL, SerQL, SPARQL. Being the last one the most promising for reasoning capabilities and widespread use.

Task 3.5 Standards-based uniform access to external sources

A solution based on uniform interfaces and existing standards is the objective of this task to enable that INTEGRATE tools and services can access data and knowledge from external repositories. Structured data sources can be queried through using the adopted standards, while unstructured datasets should be transformed into, a structured format before. Even with structured data such as EHRs, some information is still stored as free text. Simple transformations can be performed using ETLs capabilities. EHRs from the clinical partners are at the moment, the exclusive external source of the platform.

1.3.3 Deviations from the DOW and corrective actions

Expected delays to obtain surrogate data from the clinical side have caused that the WP3 have been mainly focused on Task 3.1 and 3.2. Some of the WP4 work load has been moved to the WP3 as well, since previous work on the core dataset and common data models is required to provide tools enabling data and knowledge sharing (Task 4.2).

1.3.4 Planning next period

Results from the analysis of current technologies will be used, during the next technical meeting (22nd September) and workshop (11th -12th October), to decide the technologies and models that will be used within the semantic solution (figure below).

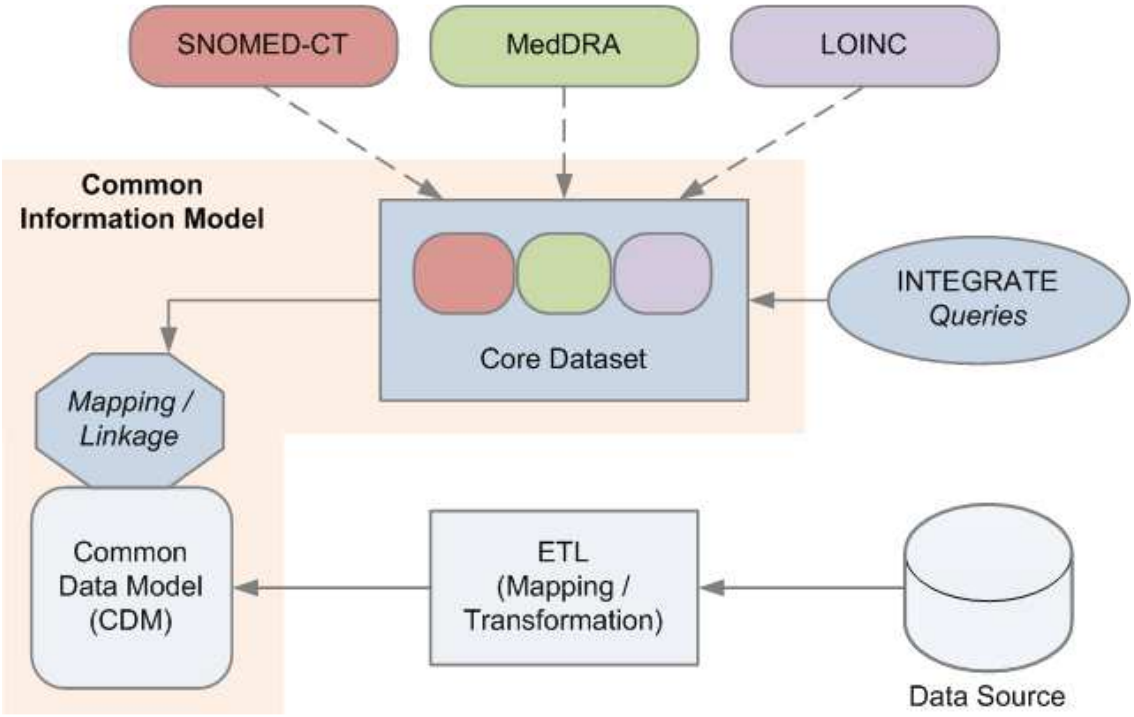

Figure 1 First approach of the semantic solution within the INTEGRATE project

The core dataset will be loaded with concepts manually identified by users and automatically identified from case report forms and eligibility criteria. A common data model will be designed and implemented, based on available solutions. And it will be populated with EHR surrogate data from the IJB partner through an ETL tool.

As shown in the next figure, at the end of the next reporting period will have a design of the common data model (D3.1 month 12) according to the INTEGRATE data requirements.

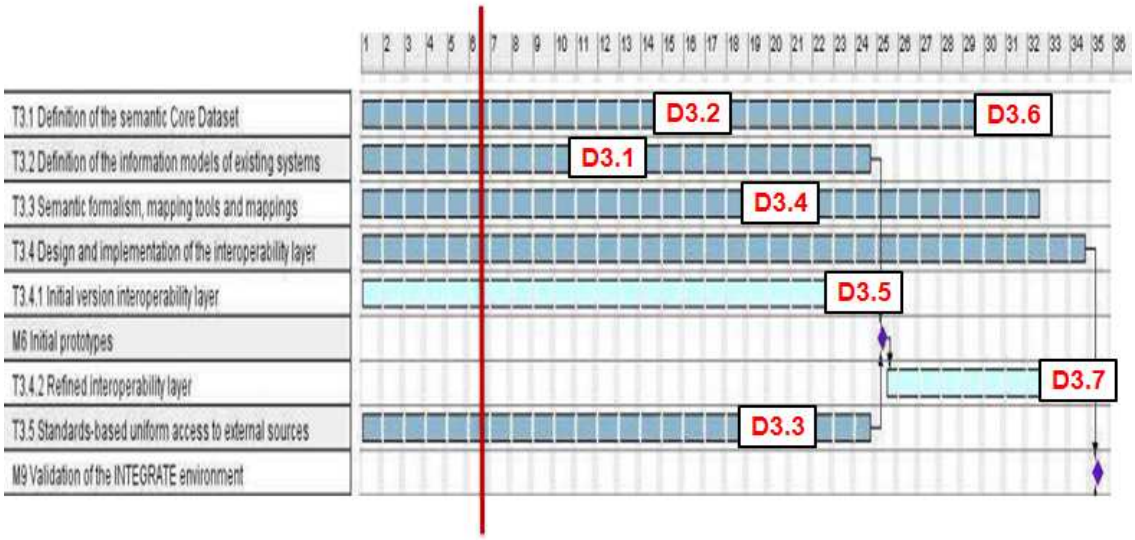

Figure 2 Overview of planning

Although the first release of the core dataset is planned beyond the next period (D3.2 month 16), we plan to provide a previous and internal version before month 12. Similarly, a first prototype of the mapping formalism and the semantic interoperability layer is expected by the 3rd INTEGRATE workshop (February 2012). The main risk to fulfill this roadmap would be an important delay on gaining access to actual data due to legal and political reasons. User requirements are highly valuable, but are unable to replace real data when developing such multi-scale integrative system.

1.4 WP4 Sharing and Collaborative Tools and Services (Lead: FORTH)

1.4.1 Objectives (of the reporting period)

The main objective of this work package is to design and develop a virtual "collaboratory" to be initially deployed and demonstrated for the BIG scientific community. Our definition and vision of scientific collaboratories is "a network-based facility and organizational entity that spans distance, supports rich and recurring human interaction oriented to a common research area, fosters contact between researchers who can be either known or unknown to each other, and provides access to data sources, artifacts and tools required to accomplish research tasks."
For this specific period a number of possible scenarios were examined regarding pathology remote collaboration concepts within BIG. Several technical discussions took place regarding the different possibilities to establish a robust collaboration framework amongst BIG participating pathologists.

1.4.2 Status/progress towards objectives WP4 (per Task)

Task 4.1 Model, data and annotation repositories
This task will develop the model library infrastructure using a common, XML-based format for the model with associated metadata description (relevant information from

3rd party data resources or literature, annotation with controlled vocabularies, results of reference analysis etc.).
Initial work on this task included the preparation of D4.1 Specification of the model, data and annotation repositories by partner Philips dealing with wide variety of data available within INTEGRATE, including clinical trial data, imaging studies, molecular (genetic) data and clinical care data, providing access to high volumes of heterogeneous biomedical data at a wide variety of spatial scales. Predictive models and simulations – stored in the model repository – will exploit this wealth of (multi-scale) biomedical data to – for instance – predict therapy sensitivity for patients, and unprecedented meta-analyses can be performed across trials. In order to efficiently access data and models, metadata and annotations are stored in metadata repositories. The work in this task first deals with the INTEGRATE scenarios (involving Molecular screening, Trial meta analysis, Predictive modelling, Central Review, etc.) in combination with relevant formats, standards and guidelines to arrive at the requirements for the data repositories, (predictive) model repositories and annotation repositories.

Task 4.2 Tools enabling data and knowledge sharing
This task will be focused on delivering a set of services and tools of the virtual collaboratory of the BIG community exploiting innovative community annotation, crowd-sourcing and scientific accreditation tools as well as semantic approaches to interoperability and automated reasoning. In addition, in order to support clinical research, the task will develop tools enabling the clinical research community to collaboratively define research protocols and carry out all the necessary regulatory and administrative steps to set up a clinical trial.
FORTH provided guidance in the technical discussion concerning the slide scanner that will be acquired by BIG, taking into account the collaboration parameters that the workflow between pathologists must have.
The list of the slide scanners of interest is the following:

- Roche (Scanner: iScanCoreo, Software: Virtuoso Digital Pathology Application Software)
- Aperio (Scanner was not specified)+Definiens Software
- Hamamatsu (Hamamatsu HPF-Nanozoomer RS2.0 PACK)
- Leica (scanner SCN400)
- Olympus (scanner & software were not specified)

In avaluating the Aperio platform, a webinar from Aperio was carefully studied. The first evaluation is that the software seems capable and expandable but the most suitable to determine that would be the Clinicians themselves.

Task 4.3 Tools enabling collaboration
An important requirement for emergent collaborations is a shared workspace that is accessible to all collaborators. Ideally, this workspace should include all the important transactions that have taken place among scientific workers. In addition to helping a group of collaborators learn from past transactions and take the best step forward, the workspace will facilitates stigmergy, i.e., it will enable a worker's contribution to stimulate others to build on that contribution without any direct communication between the workers.
FORTH initially suggested to create a central imaging review tool for BIG trials. Eventually BIG suggested to drive the development of a collaboration environment for pathologists instead.

In the webinar that was organized on the 18th of July, from Pixcelldata an interesting software platform was presented, Collibio. Subsequently, a project-internal discussion between FORTH and BIG took place. Based on the details of the presentation and the discussion that took place, we have concluded that the pros and cons of the specific platform are:

Pros

- The main concept of the platform is USERS and IMAGE SERVERS, which are brought together in PROJECTS or WORKSPACES. This appears to be an interesting approach to support collaboration.
- Multiplatform: A web based environment, accessible from any type of desktop operating system (because it is flash based, mobile operating systems are excluded, except android).
- A Remote pathology viewer: Images are not downloaded locally, but are accessed directly from the database of the Image Server of the slide scanner (it has to be supplied from the manufacturer itself). Because the images are not downloaded locally the user does not have to wait for them (which might be very time consuming).
- The images are available as soon as they are scanned.
- Support for users with configurable roles and permissions that can share projects contacting multiple images.
- Collaboration capabilities for asynchronous reporting.
- It has a very configurable mechanism to create custom forms.
- Supports navigation modes (zoom in and out) and Annotate mode.
- In general it seems to be a modular platform, providing APIs for image access and upload.

Cons

- The database which stores the links between the images and the metadata information (annotations, forms, etc) is handled by Pixcelldata, in their farm.
- It does not support a centralized image repository.
- It does not support all the major Image scanners (e.g Roche bio-imaging platform is NOT supported).
- It does not have any predefined form or pathology. The forms are generic, and cannot be imported or distribute the template of a form between the users.
- Does not support SYNCHRONOUS COLLABORATION, i.e. multiple users collaborating and interacting on the same image simultaneously (although many users can open the image at the same time, their number and the efficiency of operation is set by the total amount of users accessing the image server and the hardware capabilities of the image server. In case that the collaborators try to do a synchronous operation, database hierarchy and logic is applied and the first user gains the lock, while the changes from the rest users are rejected until the lock is released).
- The segmentation and annotation functionality of the image viewer is limited.
- The export formats are limited to only one (EXCEL). Lack of XML export is crucial for collaboration with other toolkits. Some XML export functionality is planned, but we do not know when it will be available.

- Image analysis functionality is not supported.

Based on the experience FORTH suggested a number of alternatives.
Alternative A
A commercial off-the-shelf solution is selected to fulfill the collaboration needs in INTEGRATE – to whatever degree available platforms allow.
Alternative B
FORTH undertakes the responsibility to coordinate the effort (in the context of WP4) to develop a solution that encapsulates as much functionality of the investigated products as is considered necessary, in order to create a customized, INTEGRATE specific, collaboration environment that is able to handle BIG's requirement for a centralized data (including pathology image) warehouse. Such an approach would enable us to integrate additional functionalities including image analysis (e.g. for cross-image intensity normalization, estimation etc.) and support for synchronous collaboration. FORTH has the resources required to provide such a solution that will be tailored to the specific needs of BIG and the INTEGRATE project. FORTH can also commit to provide the necessary technical support of this dedicated collaboration and analysis platform, even after the end of the INTEGRATE project.

Task 4.1 Privacy Enhancing Processes and Services
Initial work in this task is carried out by partner CUSTODIX aiming to ensure privacy and security within the specific architecture of INTEGRATE.

1.4.3 Deviations from the DOW and corrective actions
No major deviations timewise. On the other hand the collaboration concepts where somehow different between the conceptual level presented in the DoW and the actual needs of BIG. However, FORTH has been trying to balance both in the discussions regarding the collaboration environment that needs to be established within BIG pathology central review.

1.4.4 Planning next period
Define the final user requirements and needs as well as concrete use-cases for pathology collaboration tools and services and initiate the development process. This will be part of the decisions that will take place in the next plenary meeting in October at FORTH.

1.5 WP5 Support for predictive modeling and simulators (Lead: FORTH)

1.5.1 Objectives (of the reporting period)
The main objectives of this work package are to propose an approach and a methodology and to build a framework enabling the development multi-scale predictive models of response to therapy in breast cancer, making use of multi-level heterogeneous data provided by clinical trials in the neo-adjuvant setting. The models developed in this WP will be based on realistic clinical research scenarios, as outlined in the neoBIG research program, and on comprehensive data sets from rigorously conducted clinical trials. The models will also be used to validate the INTEGRATE approach and the appropriateness of the INTEGRATE infrastructure.

The main objective for this reporting period was to propose a methodology for predictive models development within clinical trials for more efficient development and validation of such models and faster adoption into clinical practice.

1.5.2 Status/progress towards objectives WP

Task 5.1 Definition of clinical scenario (questions) for the INTEGRATE VPH use case

This task uses as input the clinical scenarios elaborated in WP1, based on which will develop VPH-focused scenarios. This Task takes input of Task1.2, Task3.4-5 in order to exploit the possibilities of sharing data both provided by our clinical partners within the INTEGRATE environment but also from public databases. After several discussions data from the TOP trial are gradually becoming available for developing novel models within this WP.
Three different scenarios have been defined so far and will be used for VPH modelling development have been defined and will be reported in detail in D5.1

Task 5.2 Definition of genetic and imaging biomarkers and of a modelling methodology

The consortium decided to share the BIG data from the TOP trial in order to develop novel predictive models and investigate new biomarkers.

Task 5.3 Development of predictive models of response to therapy and of the modelling framework

A number of actions have been taken in order to define the more efficient methods for building a prediction model from different data sources. A number of approaches has been studied for different modelling aspects, including:

- Feature selection methods for selecting a subset of relevant features.
- Data integration methods for constructing an informative meta-dataset.
- Building accurate classifiers for the prediction work.
- Pattern recognition methods for estimating the generalization error of the prediction model.
- Statistical methods for evaluating the performance of the prediction model.

Based on this analysis it was decided to follow novel modelling approaches that integrate heterogeneous data. To overcome the limitations of traditional methods, a multiple kernel framework is proposed for this task, that will use a set of kernels, instead of a single one. This novel technique for combining heterogeneous information from various data sources in a common kernel framework is the Multiple Kernel Learning (MKL), pioneered by [G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. Jordan. Learning the kernel matrix with semi-definite programming. Journal of Machine Learning Research, 5, 2004.] to incorporate multiple kernels in classification. The essence of MKL relies on the kernel representation while the heterogeneities of data sources are resolved by transforming the different data sources into kernel matrices. MKL involves first transforming each data source (e.g. clinical, microarray and proteomic data) in a common kernel framework, followed by weighted combination of the individual kernels as given by the following equation. A detailed description will be given in the upcoming deliverable D5.1.

1.5.3 Deviations from the DOW and corrective actions

No deviations for this reporting period

1.5.4 Planning next period

D5.1 is under preparation and will be submitted on time. This will set the basis for all the future work in WP5.

1.6 WP6 Pilots, Evaluation and Validation (Leader: Philips)

1.6.1 Objectives (of the reporting period)

The main objectives of the work package are:

- To formulate evaluation criteria, validation procedures, and feedback report guidelines
- To coordinate the specifications of test (validation) cases and demonstrators
- To coordinate evaluation and validation activities concerning all the project software components
- To coordinate the efforts with the pilot sites
- To prepare the technical and procedural infrastructure for the installation of the INTEGRATE software solutions

The objectives for WP6 during this reporting period were to coordinate the efforts with the technical staff and the IT departments of the pilot sites, so that the Consortium receives all information required for developing the information models of the existing infrastructures, and all the data necessary for the testing and validation of the INTEGRATE infrastructure components and tools.

1.6.2 Status/progress towards objectives WP6 (per Task)

T6.1: Building the INTEGRATE development and testing environment

The objective of this task is to coordinate all efforts that need to take place locally at each and every pilot site:

- early enough in the project implementation period – to build the development environment (e.g. “surrogate” databases), and
- provide access to suitable schema- and instance-level datasets to be used by the project prototypes.

In this reporting period, the focus was to build all the necessary knowledge concerning available relevant systems, clinical workflows, data, etc., and to follow all the necessary legal and ethical steps so that the consortium can obtain access to the relevant knowledge and data.

We have evaluated the INTEGRATE needs with respect to the development environment for each workpackage and task. Based on this evaluation we have defined comprehensive requirements concerning the development and testing environment and agreed on a process (clear steps) leading to the release of the information and data by the clinical partners in compliance with legal and ethical regulations and in time to support the development of the technical solutions.

The main information and data needs identified by INTEGRATE are:

- Interfaces of relevant systems in care and research
- Schemas of relevant systems, all representative formats
- Interactions among the systems

- Data flow
- Deployment of real systems (e.g. EHR) – software
- Info about how open those systems are:
    - access to full interface,
    - can we deploy the software,
    - can we change the software
- Information on all types of data, sizes, volume
- Information on who is responsible for the data (e.g. lab), who owns the data, who enters data, where, how\who exports it
- Location of the data (now and in the future)
- Uses of the various data types
- Meta data (description data)
- Querying scenarios
- Examples of relevant analyses based on the data
- Comprehensive set of relevant concepts
- Realistic representative dataset covering the uses, semantics, ranges of data

Many of these aspects have been clarified and elaborated upon in draft deliverables concerning the specification of the INTEGRATE repositories, the clinical scenarios, and the use cases. Data of a completed trial (TOP) has been provided to the consortium following the approval of the Ethics Committee of IJB.

1.6.3 Deviations from the DOW and corrective actions

There are no major deviations from the DOW.

1.6.4 Planning next period

In the next period the effort in WP6 will concentrate on creating the right conditions for the other work packages to carry out their work, by coordinating the setting up of the development environment of the INTEGRATE project and the provision of test data (for prototype development) in compliance with all legal and ethical requirements. These efforts and the results will be the topic of deliverable D6.1 Report on the development environment and on the available test data, which is due in month 9.

1.7 WP7 Knowledge Management (Lead: BIG)

1.7.1 Objectives (of the reporting period)

The objectives for WP7 during this reporting period were to establish mechanisms of information exchange between the project members and to start giving visibility to the project through initial dissemination activities.

1.7.2 Status/progress towards objectives WP7 (per Task)

Task 7.1 Dissemination

Media for the internal dissemination of the information between the members of the INTEGRATE consortium have been established. These include a mailing list, a

devoted file-sharing server and a wiki. The members of the consortium now extensively use these media.

The main channel of information external dissemination for the INTEGRATE project is the public website (http://www.fp7-integrate.eu) which already contains information on the objectives, the strategy and brief details about the partners of the INTEGRATE project. The public website uses a content management system, which makes updates more convenient and will thus help to keep the content up-to-date.

Conferences are also seen as important dissemination channels. Two abstracts for poster presentations about the INTEGRATE platform have already been accepted in international oncology conferences. These posters will present INTEGRATE to the community of oncologists and translational researchers (see in the dissemination section).

Finally, two popularization articles describing INTEGRATE have also been accepted for publication, which will increase visibility of the project in the general public (see also in the dissemination section).

1.7.3 Deviations from the DOW and corrective actions

There are no major deviations from the DOW.

1.7.4 Planning next period

As more results and publishable material become available, BIG will ensure the update of the website on a regular basis.

Work on the initial dissemination plan has started and leverages the extensive experience of BIG with respect to dissemination activities. As part of this dissemination plan, target audiences and future dissemination events and channels will be defined.

Work on the exploitation plan for the INTEGRATE platform will also be performed during the next period.

2 Consortium management

2.1 Consortium management tasks and achievements

During this period the main tasks in WP8 were to set up a coherent way of working in the consortium and to enable effective collaboration. A first consortium workshop and a meeting with a member of the External Advisory Board have also been organized. Additionally, the setting up of relevant collaboration with prominent external initiatives has been supported.

2.2 Changes in the consortium

No changes in the consortium

2.3 Cooperation

In this first period the cooperation in the consortium has been excellent, the regular meeting and t-cons enabling a clear definition of tasks and close collaboration on the development of the technical solutions. The scenarios, use cases, the architecture and the approach towards the semantic solution are all the result of collective effort.
Additionally, we have set up relevant external collaborations with prominent initiatives in our area of research:

- SAGE Bionetworks
- TRANSCEND system / UCSF

Within these collaborations we have agreed to share clinical scenarios, requirements, and solutions.

2.4 Meetings

Date	Event	Venue/host	Country
02-03/02/2011	Kick-Off Meeting	Brussels/BIG	Belgium
23/02/2011	WP2 Meeting	Eindhoven/Philips	Netherlands
04/03/2011	Monthly Telco	Sint-Martens-Latem/Custodix	Belgium
22/03/2011	WP1 Meeting	Brussels/BIG	Belgium
01/04/2011	Monthly Telco	Sint-Martens-Latem/Custodix	Belgium
08/04/2011	WP2 Meeting	Brussels/BIG	Belgium
06/05/2011	Monthly Telco	Sint-Martens-Latem/Custodix	Belgium
09/05/2011	Technical Meeting	Brussels/BIG	Belgium
12/05/2011	Technical Meeting	Eindhoven/Philips	Netherlands
25/05/2011	Technical Meeting	Sint-Martens-Latem/Custodix	Belgium
27/05/2001	Follow-up meeting	Brussels/BIG	Belgium
03/06/2011	Legal Meeting	Sint-Martens-	Belgium

		Latem/Custodix	
09/06/2011	Legal Meeting	Brussels/BIG	Belgium
10/06/2011	Monthly Telco	Sint-Martens-Latem/Custodix	Belgium
20/06/2011	WP3 Semantics Meeting	Brussels/SOST	Belgium
21-22/06/2011	Consortium Meeting	Brussels/BIG	Belgium
22/6/2011	Meeting EA John Huffman, Poiesis Informatics	Brussels	Belgium
23/06/2011	WP2 architectural Meeting	Sint-Martens-Latem/Custodix	Belgium
01/07/2011	Monthly Telco	Sint-Martens-Latem/Custodix	Belgium
13/07/2011	WP4 Semantic Interoperability Meeting	Amsterdam/Philips	Belgium
21/4/2011	BIG EB	Teleconference, BIG	-
5/12/2011	BIG EB	Teleconference, BIG	-
6/2/2011	BIG EB	Chicago Hilton	USA
6/3/2011	BIG AC	Chicago Hilton	USA
7/7/2011	Meeting with David Cameron (Edinburgh)	Teleconference, BIG	-
9/8/2011	BIG EB	Teleconference, BIG	-
2/3/2011	IJB-BIG-Philips meeting	IJB, Brussels	Belgium
22-23/3/2011	INTEGRATE consortium meeting	IJB, Brussels	Belgium
10/4/2001	Meeting with Dr Flamen (Bordet radiologist)	IJB, Brussels	Belgium
14/4/2011	Meeting with Sarah Davis (TRANSCEND)	University of California	USA
15-16/4/2011	SAGE Meeting	University of California	USA
10/5/2011	Meeting with Dr Lemort (Bordet radiologist)		
5/6/2011	Meeting with Stephen Friend (SAGE)	IMPAKT conference, The Square, Brussels	Belgium

3 Achievements per individual partner

Partner 1 Philips

- Participated to / organized WP1/WP2/WP3/WP4 and consortium meetings
- Collaborated on the semantic solution, selection of ontologies, and definition of the core dataset. We have carried out studies and extracted relevant sets of concepts specific for breast cancer and compared to other diseases, based on a large corpus of data (all eligibility criteria of cancer trials on ClinicalTrials.gov)
- Conducting a state-of-the-art survey of semantic Web technologies and standards relevant for different aspects of the solution envisioned in the project. This effort is part of the work package 2 and the results are expected to be reported on a project deliverable due October 2011.
- In the "User Needs and requirements" work package, we have discussed and reviewed the user scenarios iteratively and extensively (leading to the D1.1 and D1.2). The activities consisted of meetings, telco's and reviewing.
- In the "Architecture and integration" work package, we have contributed to the development of the technical use cases based on the user scenarios.
- We have investigated the i2b2 system. For this purpose, an example patient case was modelled (a breast cancer patient with a medical history and various tests performed) in the i2b2 model to assess whether/how i2b2 can contribute to the INTEGRATE platform.
- In the "Sharing and collaborative tools and services" we work on the specifications of the model, data and annotation repositories. Given the user scenario's, a description was made of the (formats of the) data that should enter the INTEGRATE platform, and logical models are being developed to specify the contents of the various repositories.
- Contributed to several publications (papers and posters) and deliverables.

Partner 2 BIG

During the last period, BIG, in close collaboration with IJB (most working reunions have been joint working reunions), has achieved the following:

- Gather a panel of experts from various fields (oncologists, translational researchers, data analysts, clinical trial managers, IT) to contribute to the user requirement analysis
- Liaise with representatives of related data sharing initiative and organize meetings with them
- Gather initial information about legal, regulatory and IPR aspects
- Lift the legal hurdles for reuse of data from previous projects and clinical trials within INTEGRATE
- Setup the external website in collaboration with FORTH and populate it with information about the INTEGRATE project.
- Contribute to articles and accepted conference abstracts for the dissemination of the project.

Partner 3 FORTH

- Co-development of project's website
- Development of project's wiki
- Participated on several discussions with BIG and elaborated a document evaluating the technological possibilities discussed for establishing a collaborative environment for pathology
- Started working on D5.1 for predictive modelling development and proposed novel mathematical concepts that can be implemented within this WP and which give the potential for much more sophisticated model development.

Partner 4 CUSTODIX

- Attended WP1/WP2/WP3/WP4 and consortium meetings
- Created first draft document for use cases
- Created first draft component model
- Created first attempt security model
- Contributed security section of the state-of-the-art document
- Researched re-usable/available relevant solutions and components (like I2B2)
- Discussed and provided input for the user requirements and scenarios of WP1
- Initial architecture brainstorm sessions
- Contribution in discussions about semantic approaches, data sources and common and local information models

Partner 5 IJB

During the last reporting period, IJB, in close collaboration with BIG (most working reunions have been joint reunions), have achieved the following:

- Define the basic functions and the categories of users of the platform
- Define and draft the description of the user scenarios
- Contribute to the definition of the list of semantic core concepts
- Gather and share with the other partners of INTEGRATE preliminary clinico-genomic data
- Perform a preliminary evaluation of some reusable software components (e.g. caTissue for biotracking)
- Investigate, in collaboration with FORTH and the IJB pathology department, several solutions for central review of pathology
- Test conversion of information from the IJB electronic health records to an interoperable format

Partner 6 UPM

- User requirements understanding
- Involved within the project communications mechanisms (periodic teleconferences, physical meetings, etc.)
- Collaboration to provide architectural principles and design
- Analysis of core dataset candidates
- Analysis of core dataset formats

- Collaboration to analyse available data models to design the INTEGRATE common data model
- Analysis of Extract Transform and Load Tools
- Collaboration to design the semantic solution of the INTEGRATE platform

4 Deliverables

Deliverable no.	Deliverable name	Lead Partner Acronym	Work Package	Due Date (DoW)	Actual Delivery Date
				Project Month	
D8.1	Public summary of the project	Philips	WP8	1	1
D7.1	Communication portal Wiki	BIG	WP7	3	3
D8.2	Internal project website	Philips	WP8	3	3
D1.1	User needs and specifications	IJB	WP1	6	6
D7.2	External project website	BIG	WP7	6	6

4.1 List of milestones

Milestone no.	Milestone name	WP no.	Due date (DoW)	Actual date	Lead partner
MS1	Formation of boards and committees	WP7	6	3	BIG
MS2	Initial requirements w.r.t. the INTEGRATE environment	WP1	6	6	IJB

5 Use and dissemination

5.1 Dissemination activities

Planned /actual Dates	Type	Type of audience	Countries addressed	Size of audience	Partner responsible /involved
Sep 23-27, 2011	Poster in international oncology conference	Medical, radiation, and surgical oncologists, translational researchers, basic scientists, healthcare workers, patient advocates,	All European countries, and also strong representation from most other countries	Estimated 15000	BIG, Philips
Nov 3-5, 2011	Poster in international oncology conference	Medical, radiation, and surgical oncologists, translational researchers, basic scientists, healthcare workers, patient advocates,	European as well as other continents	Estimated 2500	BIG, Philips
Dec, 2011	Article in newsletter	Clinical trialists, Medical, radiation, and surgical oncologists, translational researchers, basic scientists, healthcare workers, patient advocates,	European as well as other continents	50 clinical trial groups worldwide	BIG
Q4, 2011	Article in magazine	General public	Mainly Belgium		BIG

5.2 Publications

Two articles describing the INTEGRATE platform have been approved for publication and will be published before the end of 2011:

- The first article will feature in the newsletter of the Breast International Group, volume 13/2, which will be published in December 2011.
- The second article will feature in the “Subsidies” magazine edited by the Belgian National Lottery and presents the projects that they are funding.

5.3 Contributions to conferences (abstracts, etc)

Two abstracts describing the INTEGRATE platform have been accepted for presentation as posters during these international oncology conferences:

- the ECCO-ESMO-ESTRO multi-disciplinary cancer conference, September 23-27 2011, Stockholm, Sweden
- the ABC1 Advanced Breast Cancer Conference, November 3-5 2011, Lisbon, Portugal

Three publications have been accepted in international conferences in healthcare informatics and knowledge management:

- " Classification of clinical trial eligibility criteria to support semantic linkage of research and clinical care data ". AMIA 2011 Annual Symposium October 22-26, 2011
- "Patterns of clinical trial eligibility criteria"
  The 3th International Workshop on Knowledge Representation for Health-Care 2011, 06/07/2011
- "Patterns of clinical trial eligibility criteria", Compressed contribution, 23rd Benelux Conference on Artificial Intelligence, 03/11/2011

6 Manpower overview

Actually Spent 6-Monthly Human Resource Allocation

Partner	WP1		WP2		WP3		WP4		WP5		WP6		WP7		WP8		*Total*	
	planned	spent	Planned	spent	Planned	spent	Planned	spent	planned	spent	planned	spent	Planned	spent	planned	spent	planned	spent
Philips	0.6	1.0	1.6	2.5	1.7	3.0	1.9	0.5	0.6	0.0	0.9	0.5	0.3	0.5	1.0	1.0	8.6	*9.0*
BIG	1.5	1.5	1.1	0.6	0.1	0.1	0.9	0.9	1.0	0.6	0.8	0.8	0.9	0.9	0.12	0.12	6.42	*5.52*
FORTH	0.6	0	0	0	0.5	0	2.3	2.3	2.7	3	0.8	0	0.2	0.37	0.25	0	9.35	*5.67*
Custodix	0.5	3.18	3.5	5.80	0.5	0.11	1.7	0	0.2	0	0.7	0	0.1	0	0.08	0.05	7.28	*9.14*
IJB	2.3	3.5	0	0	0.6	1	0	0	1.2	1.2	0.9	0.9	0.3	0.3	0.12	0.12	5.42	*7.02*
UPM	0.5	0.54	1.5	1.63	2.3	3.6	1.2	0.75	0.2	0.12	0.8	0.6	0.2	0.19	0.12	0.15	6.82	*7.58*
Total WP	*12.0*	*9.72*	*19.4*	*10.53*	*11.4*	*7.81*	*16.0*	*4,35*	*11.8*	*4.92*	*9.8*	*2.8*	*4.0*	*2.26*	*3..34*	*1.44*	*87.8*	*53.65*

(actual man months are 6-monthly best estimates; final accurate man-hours are given in the cost claims)