Integrative Cancer Research
INTEGRATE
Through Innovative Biomedical Infrastructures

SEVENTH FRAMEWORK PROGRAMME

ICT-2010-270253

INTEGRATE

Driving excellence in Integrative Cancer Research through Innovative Biomedical Infrastructures

STREP
Contract Nr: 270253

Deliverable: 6.1 Report on the development environment and on the available test data

Due date of deliverable: (10-31-2011)
Actual submission date: (12-16--2011)

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

Responsible WP: Philips

Revision: final

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		
Dissemination level		
PU	Public	x
PP	Restricted to other programme participants (including the Commission Service	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (excluding the Commission Services)	

0 DOCUMENT INFO

0.1 Author

Author	Company	E-mail
Anca Bucur	Philips	anca.bucur@philips.com
Lina Pugliano	BIG	lina.pugliano@bordet.be
Philippe Hennebert	IJB	philippe.hennebert@bordet.be
Jasper van Leeuwen	Philips	jasper.van.leeuwen@philips.com
Mohammed Jelti	IJB	mohammed.jelti@bordet.be
Alexandre Irrthum	BIG	alexandre.irrthum@bordet.be

0.2 Documents history

Document version #	Date	Change
V0.1	11/24/2011	Starting version, template
V0.2	11/24/2011	Definition of ToC
V0.3	12/05/2011	First complete draft
V0.4	12/05/2011	Integrated version (send to WP members)
V0.5	12/06/2011	Updated version (send PCP)
V0.6	12/08/2011	Updated version (send to project internal reviewers)
Sign off	12/16/2011	Signed off version (for approval to PMT members)
V1.0	12/16/2011	Approved Version to be submitted to EU

0.3 Document data

Keywords	
Editor Address data	Name: Anca Bucur Partner: Philips Address: High Tech Campus 34, (6-030) 5656AE Eindhoven, The Netherlands Phone: +31 40 27 49709 Fax: E-mail: anca.bucur@philips.com
Delivery date	

0.4 Distribution list

Date	Issue	E-mailer
12/06/2011	V0.6	fp7-integrate@listas.fi.upm.es
12/16/2011	V1.0	Joel.Bacquet@ec.europa.eu

Table of Contents

1 Introduction

An important task of WP6 is to build the development environment and create the set-up allowing the INTEGRATE infrastructure components and tools to be designed and built. This includes providing sufficient schema-level data describing the structure and content of both the clinical and research infrastructures, to enable the development of canonical information models for the electronic health record (EHR) and clinical trial (CT) systems. Examples of such schema-level data are case report form (CRF) fields from clinical trials, and archetypes of EHR data. Additionally, sufficient (and well-matched) patient data from both care and research systems need to be provided to enable the testing and validation of our solutions. This data needs to be prepared according to the legal, privacy and security requirements. All this data will be stored in “surrogate” information systems that will be used during development, in order not to disrupt the clinical environment.

We need 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.

As a first step we have built all the necessary knowledge concerning available relevant systems, clinical workflows, data, etc., and we are following 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 so far 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
- Information 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 the deliverables concerning the specification of the INTEGRATE repositories, the clinical scenarios, and the use cases. Other aspects will require further investigation. This report focuses on describing the datasets available for the development environment of INTEGRATE and on the steps taken to generate the datasets and to comply with the relevant legal and ethical frameworks.

1.1 Aim of this Document and Outline

In the first period of the project the effort in WP6 has concentrated 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. Next, to access to data, access to and thorough understanding of the relevant systems that need to be integrated in our environment is crucial. These efforts and their results are the topic of this deliverable.

Section 2 will describe the TOP trial, a completed study whose data will be available to integrate. In Section 3 we will describe the Pilot Study designed to support the implementation of the molecular screening service. Section 4 will describe the context of the EHR system and the patient data that will be available for the development of the INTEGRATE semantic layer. Section 5 will present the main steps required to obtain the necessary approvals from the Ethics Committees for the use of patient data in INTEGRATE. Our conclusions will summarize the progress achieved so far with respect to getting access to relevant datasets and the next steps.

2 The TOP Trial

2.1 Overview

The Trial of Principle (TOP)[1] study[2] was started for the prospective evaluation of Topoisomerase II Alpha (*TOP2A*) gene amplification and protein overexpression as markers predicting the efficacy of Epirubicin in the primary treatment of breast cancer patients. The TOP trial included 149 patients, 139 of whom were evaluable for response prediction analyses. The primary end point was pathologic complete response (pCR). TOP2A and gene expression profiles were evaluated using pre-epirubicin biopsies. Gene expression data from ER-negative samples of the EORTC (European Organisation for Research and Treatment of Cancer) 10994/BIG (Breast International Group) 00-01 and MDACC (MD Anderson Cancer Center) 2003-0321 neoadjuvant trials were used for validation purposes.

The following inclusion and exclusion criteria were defined:
Inclusion Criteria:

- Histologically-confirmed breast cancer (operable, locally advanced or inflammatory)
- Age less than 70 years
- Female patient
- Tumor size 2 cm at ultrasound examination.
- ER-negative tumors defined according to immunohistochemistry (i.e. < 10% of positive cells after immunostaining).
- Multifocal and multicentric breast tumors are allowed if all foci are ER-negative.
- Fixed and frozen samples from the primary tumor, obtained before treatment with epirubicin, must be available for evaluation of biological markers (TOP2A gene and protein, HER-2 gene, p53 gene, oligonucleotides microarrays).
- Written informed consent before study registration.
- Performance status 0 or 1 (ECOG scale)
- Normal CBC, hepatic and renal functions
- Normal left ventricular ejection fraction by echocardiography or muga scan
- Negative pregnancy test for all women of childbearing potential. Patients of childbearing potential must implement adequate non-hormonal measures to avoid pregnancy during treatment.

Exclusion Criteria:

- Metastatic breast cancer
- Serious medical conditions like:
    - congestive heart failure or unstable angina pectoris, previous history of myocardial infarction within 1 year from study entry, uncontrolled arrhythmias.
    - history of significant neurologic or psychiatric disorders
    - active uncontrolled infection

[1] Desmedt C, Di Leo A, de Azambuja E, Larsimont D, Haibe-Kains B, Selleslags J, Delaloge S, Duhem C, Kains JP, Carly B, Maerevoet M, Vindevoghel A, Rouas G, Lallemand F, Durbecq V, Cardoso F, Salgado R, Rovere R, Bontempi G, Michiels S, Buyse M, Nogaret JM, Qi Y, Symmans F, Pusztai L, D'Hondt V, Piccart-Gebhart M, Sotiriou C. Multifactorial approach to predicting resistance to anthracyclines. J Clin Oncol. 2011 Apr 20;29(12):1578-86

[2] http://clinicaltrials.gov/ct2/show/NCT00162812?term=top&rank=6

    - active peptic ulcer, unstable diabetes mellitus
- Concomitant contralateral invasive breast cancer
- Concurrent treatment with hormonal replacement therapy
- Concurrent treatment with any other anti-cancer therapy
- Previous treatment with anthracyclines for breast cancer

And the schedule of assessments is documented in **Table 1**.

Mandatory Exams	**Baseline** **< 28 days before 1st infusion**	**Epirubicin Treatment period**	**Post-Epirubicin Treatment**
Medical history	X		
Physical examination + clinical tumor assessment	X	X	X
Breast biopsy (TRU-CUT) + measurement of hormone receptors	X	(X)	
Serum Sample	X	X	X
Whole Blood Sample	X		
Hematology and Biochemistry Red blood cells Hemoglobin Platelets WBC ANC Total Bilirubin Serum Creatinine GOT/GPT Alkaline Phosphatase	X	X	X
ECG	X		
LVEF (US or MUGA)	X		
Chest X-Ray	X		
Bone Scan	X		
Liver Ultrasound	X		
Bilateral Mammography	X		X
Breast Ultrasound	X		X
Informed consent	X		

Table 1 - schedule of assessments

2.2 Available dataset

There is a wide variety of data available from the TOP trial. The available data comprises the Case Report Forms, aggregated clinical data, and genomic data (including gene expression microarray data, SNP data and methylation data). At the moment, the possibility of including imaging data is being investigated.

2.2.1 Case Report Forms

Data is collected during clinical trials from the participating sites using Case Report Forms (CRF) – paper or electronic questionnaires. The CRFs contain all data of a patient collected during a trial, including eligibility criteria and adverse events.
Oracle Clinical was used to capture the data from the CRF's for the TOP trial. The CRF data is provided to INTEGRATE in excel format and follows the structure of the Oracle Clinical database.

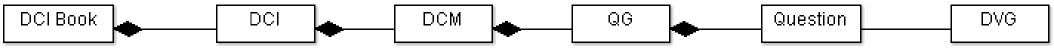
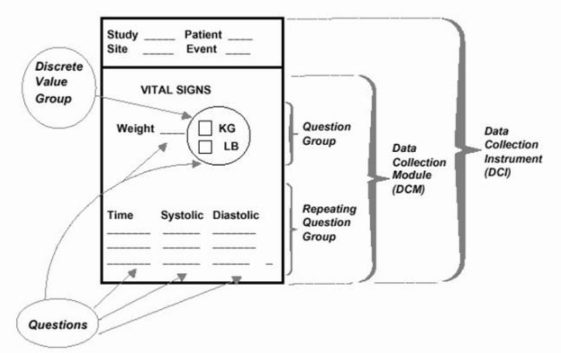


Figure 1 - Oracle Clinical structure

As documented in Deliverable 4.1 (here duplicated for sake of completeness), Figure 1 shows an overview of the structure of the Oracle Clinical database. *Question* represents a (particular) question on a CRF and its answer. When the answer can be populated from a set of possible answers (e.g. a code list), the answers are store in *DVG*, a Discrete Value Group.

Questions are grouped in *QG* – Question group – which can be used to group (medically) related questions (which can be handy in order to reuse groups of questions of different CRFs). The question groups are used in *DCM* – Data Collection Module – which represent (the sections of) the CRF screens that are used to collect the data and these sections should be answered during a single clinical visit.
DCI – Data Collection Instrument – corresponds to a CRF. Typically, 1 DCI corresponds to 1 DCM, but can include more that one DCM. The DCI construct allows for CRFs which collect data during multiple visits.
Finally, a DCI Book specifies the order of the DCI's.



The above description describes the Oracle Clinical database structure. In addition to the database structure, Oracle Clinical allows for the definition of validation and derivation procedures. Validation procedures validate data entry. Derivation procedures create additional variables in the database to assist in analysis. The prototypical example would be a derivation procedure which would calculate the age based on the birth date (as entered on a CRF) and a visit date (also entered on a CRF).

Patient's Characteristics
DCM=QG=VS New

- Height I__|__|__|__| (cm) ***HGT*** ***HGT_UN(dft 'CM')***
- Weight I__|__|___,__| (Kg) ***WGT*** ***WGT_UN (dft 'KG')***
- BSA |___,___| (m^2) ***BSA*** ***(NUMBER 4,3) BSA_UN (dft 'M2')***
- Menopausal status: ***REP_STAT(DVG REP_STAT#New)***

□$_1$ premenopausal (< 6 months since last menstrual period (LMP) and no prior ovariectomy and no estrogen replacement therapy)

□$_2$ postmenopausal (prior bilateral ovariectomy, or > 12 months since LMP with no prior hysterectomy and not receiving LH-RH analog)

□$_3$ above category not applicable and < 50

□4 above category not applicable and ≥ 50

Figure 2 - Annotated CRF excerpt

The excel export contains a sheet per DCM. **Figure 2** shows an excerpt from an annotated page of the TOP trial CRF. Data filled in this CRF would be exported onto an excel sheet named "VS New". Table 2 shows the (transposed) excel export with the questions from the CRF excerpt in bold.

Variable name	**value**	**value**	**value**
PT	XXX	XXX	XXX
STUDY	TOP	TOP	TOP
DCMSUBNM	VS	VS	VS
PATIENT_POSITION_ID	XXX	XXX	XXX
CPEVENT	SCREENING	SCREENING	SCREENING
DCMNAME	VS	VS	VS
SUBSETSN	1	1	1
DCMDATE			
DOCNUM	XXX	XXX	XXX
ACCESSTS	16-09-2004 10:35	16-09-2004 11:04	16-09-2004 13:40
LOGINTS	16-09-2004 10:35	16-09-2004 11:04	16-09-2004 13:40
LSTCHGTS	16-09-2004 10:35	02-05-2006 13:19	22-05-2006 11:17
LOCKFLAG	N	N	N
DCMTIME			
ACTEVENT	1	1	1
SUBEVENT_NUMBER	0	0	0
VISIT_NUMBER	1	1	1
QUALIFYING_VALUE	6	6	6

QUALIFYING_QUESTION	20007	20007	20007
REPEATSN	1	1	1
FIRST_BOOK_PAGE	6	6	6
RECEIVED_DCM_STATUS_CODE	PASS 1 COMPLETE	PASS 1 COMPLETE	PASS 1 COMPLETE
HGT	153	163	165
HGT_UN	CM	CM	CM
REP_STAT	PREMENOPAUSAL	PREMENOPAUSAL	POSTMENOPAUSAL
WGT	50	52	95
WGT_UN	KG	KG	KG
BSA	1.45	1.55	2.2
BSA_UN	M2	M2	M2

Table 2- Excel export

2.2.2 Aggregated clinical data

The aggregated clinical data comprise information on tumour size, auxiliary lymph node status, tumor grade, biomarker expression status (estrogen receptor, progesterone receptor, HER2, TOP2A), and several clinical endpoints such as pathological complete response, distant metastasis-free survival and overall survival.

2.2.3 Genomic data

Three kinds of genomic information are available: whole human genome expression array data, SNP data, and methylation data.

The whole human genome expression array data are Affymetrix GeneChip® Human Genome U133 Plus 2.0 Arrays[3]. This microarray contains probes for more than 38,500 transcripts corresponding to well-characterized genes and Unigene genes, giving a full-genome view of gene expression. The analysis will start from the “raw” .cel files that contain probe-level intensity data. This allows various schemes of data normalization and probeset data aggregation. The .cel files also contain the necessary information for array and hybridization quality assurance. The size of a .cel file for this microarray is around 32MB. Data is available for 120 patients from the TOP trial.

The SNP data are Affymetrix SNP 6.0 Arrays[4]. This microarray contains probes for more than 906,600 single nucleotide polymorphisms (SNPs) and more than 946,000 probes for the detection of copy number variation (CNV). This corresponds to a median inter-marker distance in the genome of less than 700 nucleotides. Again, the analysis will start from the .cel files, which allows maximum flexibility in the choice of the algorithms for CNV genotyping. The size of a .cel file for this array is around 61MB. Data is available for 70 patients from the TOP clinical trial

[3] http://www.affymetrix.com/browse/products.jsp?productId=131455&categoryId=35760#1_1

[4] http://www.affymetrix.com/browse/products.jsp?productId=131533&navMode=34000&navAction=jump&aId=productsNav#1_1

The methylation data are Illumina Infinium HumanMethylation27 BeadChips[5]. This array allows to interrogate the methylation status of 27,578 highly informative CpG sites located in the proximal promoters of 14,475 protein coding genes. This corresponds to an average of two interrogated CpGs per genes although a subset of >200 cancer-related genes have 3-20 interrogated CpGs. N.B. At the moment only processed data is available, the possibility of providing raw data is being investigated.
The Infinium assay uses a pair of probes for every CpG, with one probe measuring the level of the methylated CpG and the other probe measuring the level of the unmethylated CpG. The methylation of the CpG is then often expressed as a beta value, which is the ratio of the methylated signal on the sum of the methylated and unmethylated signal. Thus, beta values vary from 0.0 for a fully unmethylated CpG to 1.0 for a fully methylated CpG. The data is available for (at least) 34 patients from the TOP trial.

[5] http://www.illumina.com/products/infinium_humanmethylation27_beadchip_kits.ilmn

3 The Pilot Study

The pilot study aims to demonstrate the feasibility and to set in place methodologies enabling a molecular screening service for future clinical trials sponsored by the Breast International Group (BIG) through the use of the INTEGRATE environment. Therefore, the data collected and processed through the Pilot Study matches the molecular screening scenario proposed in INTEGRATE and the corresponding use cases. This provides a very realistic context for the implementation of the INTEGRATE tools and services.

In this section we will first give an overview of the Pilot Study and its role in the context of INTEGRATE. Next, we will describe the relevant steps and the data that will be collected.

3.1 Overview

The vision of INTEGRATE is to empower the researcher with the unique opportunity to access breast-cancer multi-scale data from clinical trials conducted throughout Europe and the rest of the world under the BIG umbrella. We start by focusing on completed neoadjuvant and/or adjuvant trials of BIG, but ultimately we aim to extend the use of the environment to all future BIG trials.

Our infrastructure aims to facilitate to two main needs in the current climate of clinical trials. The first aim is to enable data sharing and “meta-analytical” type research approaches by providing access to data from multiple clinical trials. Within INTEGRATE the researcher will be able to access molecular, pathological, imaging and clinical data from multiple completed breast cancer trials.

Secondly, thanks to recent advances in genomic technology, researchers have been able to identify important molecular aberrations that could be potential targets for novel drugs development. The INTEGRATE platform’s second aim will be to bring to a large number of hospitals an efficient molecular screening service from key European laboratories in order to identify in rapid manner patients suitable for future molecularly-driven BIG trials.

3.2 Molecular screening

The molecular screening service will provide the pathological and molecular data required for the eligibility criteria for BIG trials that will be executed through INTEGRATE. This service will rely on a series of core laboratories, located throughout Europe that will perform the required analysis of biological specimens. It is important to note that this service is not being developed to provide a population-based molecular screening for all breast cancer patients, but only for those pre-selected as potentially suitable to be enrolled in future BIG molecularly-defined trials.

The process is envisioned as follows:

1. A patient is reviewed at a hospital and has matched all the pre-screening eligibility criteria, and is therefore potentially eligible for a molecularly-defined BIG trial.
2. Once a biopsy is performed (formalin fixed paraffin embedded [FFPE] and frozen tissue), the clinician accesses the INTEGRATE system and enrols the patient in molecular screening process.

3. The patients' tumour samples will be sent to the appropriate laboratory(ies) based on proximity and type of assays required by the particular BIG trial.
4. The central laboratories will perform the assays and the result will be entered in the INTEGRATE platform. The clinician can then access the result and discuss this with the patient.

In order to develop the molecular screening service it is needed from the clinical side to 1/ develop the assays to CLIA- or ISO-standards and 2/ pilot the process required with three key central laboratories using prospectively collected samples. A prospective pilot phase is vital to ensure the smooth running of a molecular screening diagnostic service that can interface with the INTEGRATE platform prior to implementation in a future BIG clinical trial.

From the technology research perspective, it is necessary to have access to the data from the pilot study to be used in the development of the necessary semantic solution and in the implementation of the molecular screening service.

3.3 Pilot Molecular Screening service (on prospectively collected samples)

The screening service will be piloted in a prospective manner in order to ensure we can provide the quick turn-around timeline necessary for a prospective neoadjuvant trial and smooth running of the platform prior to the screening being implemented in a future BIG trial. The pilot will consist of 20 patients in three participating hospitals that are likely to recruit patients to future BIG trials.

Patients with newly diagnosed breast cancer will prospectively undergo biopsy at the participating hospital and the tumour samples (FFPE and frozen tissue) will be sent to the closest central testing laboratory (for example, patients enrolled into the study at IJB will have their sample sent to IJB for ER, HER2, Ki-67, affymetrix gene profiling expression and GGI).

The data collected will be used for the implementation of the molecular screening scenario of INTEGRATE.

4 The link to the Electronic Health Record

In this section we describe the EHR system that is deployed at Jules Bordet Institute and our approach to access data from this system and link to it within the INTEGRATE environment. Next to selected data that is relevant for the clinical scenarios and which will be pushed to a development data-warehouse to be used for prototype implementation and testing, we also have access to a deployment of the EHR software to build and test our interoperability approach, and also to the structure of the EHR system (subsystems, tables in the databases, interfaces, vocabularies, etc.).

Similarly to the pilot study, the anonymized data extracted from the EHR and stored in our data-warehouse will support in the first phases of the project the molecular screening scenario.

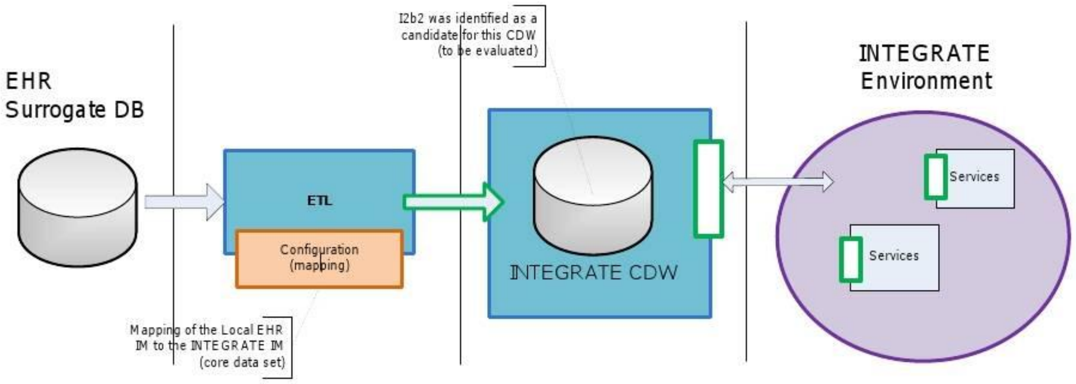

I2b2 was identified as a candidate for this CDW (to be evaluated)
EHR Surrogate DB
INTEGRATE Environment
ETL
Configuration (mapping)
INTEGRATE CDW
Services
Services
Mapping of the Local EHR IM to the INTEGRATE IM (core data set)

Figure 3 Accessing EHR data for the development environment

Figure 3 depicts our agreed approach to accessing data from the EHR.

4.1 Overview

Institut Jules Bordet (IJB) have designed and implemented their own Electronic Health Record software, Oribase. From the start the focus was on semantic interoperability norms, in particular HL7 and Clinical Document Architecture (CDA) formats in a production environment. Additionally, they have carried out work on the development of models for EHR data, collected at various key points in the patient's clinical process, that were coded and normalized using international coding systems (such as LOINC, SNOMED CT, ICD9 and ICDO) for domains relevant to screening for inclusion in a clinical trial and adverse event reporting. These choices fit very well with the standards-based approach of INTEGRATE in the development of the semantic layer.

4.2 The DB2 Database

This section briefly describes the structure of the main database (DB2) and of other resources involved in the Electronic Health Record system from IJB. We focus both on structured information (e.g. CDA documents, Laboratory Results) and on free-text reports such as radiology reports.

Demographics are stored into a central DB2 table. This table contains additional information about the patient and an identifying number called NRDO. This unique identifier refers to an instance of a record in the EHR system.

An electronic record is created for each care episode a patient receives from an ancillary department, such as radiology, laboratory, or as a result of an administrative action. The system supports the integration of healthcare data from a participating collection of systems for a single patient record. Various interfaces are needed to import data from the ancillary systems.

4.2.1 Administrative data

Admission, discharge and transfer (ADT) data are key components of the system. Several tables are involved in the ADT subsystem such as the table called Borprd/DOSS01P which stores patients' demographic data (record id, name, date of birth, gender, etc.). Table Borprd/RDMV60P contains medical appointments record. Each record contains physician id, record id, consultation type, comments, data/time. Similarly, table Bordet/HOSP03P holds hospital-stays information which includes patient record id, admission, discharge date, location, attending physician, etc. Registration regarding radiation oncology treatments is collected through an in-house program. The Borprd/RXRA79P table contains unique patient id, apparatus id, treatment period (start and end date) and the target of the treatment. Table Borprd/CHAN7WP includes surgical procedure details including the surgery type code.

4.2.2 Laboratory data

Laboratory data is integrated entirely into the system through an inbound synchronous interface. The HL7 messages sent by the laboratory system are parsed and stored in a dedicated table.

4.2.3 Clinical data

The documentation system is currently designed to use a common structure for persistent documents (CDA). It contains semi structured XML based human readable text and structured machine readable data based on LOINC and SNOMED-CT vocabularies that provide well-defined meaning for specific terms. The Bordet/BOCDS0P table stores all data for every document submitted by users. Furthermore, all user changes to a clinical document are stored as a new version of document. Each version of the document receives a record in the table.

Examples of clinical documentation that can be found in Bordet's EHR system are:

- Consultation note
- Surgery procedure report
- Echocardiography report
- Endoscopy report
- Hospital Discharge Summary report
- Radiology report

The adoption of interoperability standards, tools, architectures, and vocabularies are an important challenge for the next years. The goal is to code all documents with LOINC/SNOMED CT. This is a challenging task as some concepts are too ambiguous and other are specific to the institute. In such cases, an internal codification system is used, based on a globally unique ISO identifier attributed to IJB. These internal codes are defined hierarchically and are currently used in the coding sections of documents and other elements contained in the CDA specific to IJB.

4.2.4 Anatomical pathology data

The AnaPath system stores data in a local database and sends HL7-based information to the EHR system. All reports follow the HL7 models but contain unstructured text except for the tumor codification in SNOMED.

4.2.5 Multidisciplinary oncology consultations data

Multidisciplinary oncology consultations are group meetings between oncologists, surgeons, radiation therapists, anatomical pathologists and general physicians to discuss and decide the patient follow up. This data is stored in a central table named cmotour in the borprd collection. The data entered is structured to simplify manipulation. Internal codes are used to represent values that appear in the documents. The storage of structured data is done in the same way as of the CDA documents.

4.2.6 The cancer registry

The cancer registry is a repository containing all the cases of cancer appearing in a defined population. In this repository the characteristics of the patients are listed. It also contains the clinical and anatomopathological data collected from different sources of information. From technical point view, the cancer registry data is stored in a relational database.

5 Ethics Committee Requirements

Of prime importance for the INTEGRATE consortium is to make sure that all the data used for the development of our environment and tools is obtained and used in full compliance with all legal, regulatory and ethical requirements. In this section, we briefly describe the main requirements and steps carried out in relation with the different data sets that will be used in our development environment.

5.1 Clinical trial data

The TOP trial is a closed study whose data has been provided to the INTEGRATE project to be used for prototype development. The use of this data has been approved by the Ethics Committee of IJB, following a request from our side that explained what the project is and wants to achieve and why this data was necessary. In addition, we informed the Ethics Committee that the data that we use has been anonymized and no personal information will be accessed.

5.2 Pilot study data

Another source of data for our development environment will be provided by the pilot study. The process to obtain approval for this data is much less complicated as the primary purpose of the study is the INTEGRATE sharing environment.
The Informed Consent Form (ICF) will be specifically designed for this purpose and will inform the patients that the data will be used for the development of the INTEGRATE environment and will be stored on the platform.

5.3 Electronic Health Record data

As an important goal of INTEGRATE is to semantically link the data in our environment to the data in the Electronic Health Record System, we need access to such data for the development of the semantic layer of INTEGRATE.
Obtaining access to EHR data is much more complex both with respect to obtaining the required approvals and from logistical point of view.
First, the patients need to give informed consent for their data to be extracted/pushed onto the INTEGRATE development environment. This is likely to be covered by the initial integrate ICF for patients at the hospitals that will have an EHR link.
Next, the Ethics Committee needs to give approval per for each hospital that participates in the INTEGRATE environment. The application again needs to state what the project is, why this data is needed and what the intent of use of this data is. Approval will also need to be granted from the director of each hospital involved in the INTEGRATE that provides the EHR link. Finally, a declaration to “Commission for the protection of privacy” needs to be made.

6 Conclusions

This deliverable provides an overview of the various sources of data and information that will be provided to the INTEGRATE consortium by our clinical partners for the development of our solutions. We have also briefly described the EHR system that is currently available to be linked in our semantic interoperability layer and the approach to accessing EHR data and to integrating the EHR in our environment. During the project we will extend this work to include other relevant systems (e.g. EHRs from other hospitals, CTs).

As obtaining access to patient data is a complex task which requires strict adherence to legal and ethical frameworks, we found relevant to also describe the steps we took to obtain all the necessary approvals for the data that will be used in the project.

Next to the data provided by the consortium members, we also use in the development of our prototypes large volumes of publically available data and knowledge: ontologies such as SNOMED and MedDRA, public repositories of biomedical data, clinical trial eligibility criteria extracted from ClinicalTrials.gov, etc.

Based on the results presented in this document, we strongly believe that we have sufficient information and data to proceed with the implementation of the detailed clinical scenarios and use cases defined by the project. At the same time, we will further work to gather significant additional datasets that will enrich our development environment during the running of the project. This will enable us to support all workpackages in the project to access all data and information necessary to achieve their goals.