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: D6.4 Report on preparation of the deployment environment for the clinical pilots

Due date of deliverable: (01-31-2013)
Actual submission date: (MM-DD-YYYY)

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

Responsible WP: IJB

Revision: proposed

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
Jérôme Lhaut	IJB	jerome.lhaut@bordet.be
Cyril Krykwinski	IJB	cyril.krykwinski@bordet.be
Haridimos Kondylakis	FORTH	kondylak@ics.forth.gr
Ioannis Karatzanis	FORTH	karatzan@ics.forth.gr
George Manikis	FORTH	gmanikis@ics.forth.gr
Thanasis Iliopoulos	FORTH	ailiop@ics.forth.gr
David Pérez del Rey	UPM	dperez@infomed.dia.fi.upm.es
Raúl Alonso	UPM	ralonso@infomed.dia.fi.upm.es
Juan Manuel Moratilla	UPM	jmmoratilla@infomed.dia.fi.upm.es
Antonio Rico	UPM	arico@infomed.dia.fi.upm.es
Brecht Claerhout	Custodix	brecht.claerhout@custodix.com
Kristof De Schepper	Custodix	kristof.deschepper@custodix.com
Anca Bucur	Philips	anca.bucur@philips.com

0.2 Documents history

Document version #	Date	Change
V0.1	06/12/2012	Starting version, template
V0.2	06/12/2012	Definition of ToC
V0.3	01/03/2013	First complete draft
V0.4	15/03/2013	Integrated version (send to WP members)
V0.5	25/03/2013	Updated version (send PCP)
V0.6	01/04/2013	Updated version (send to project internal reviewers)
Sign off	05/04/2013	Signed off version (for approval to PMT members)
V1.0		Approved Version to be submitted to EU

0.3 Document data

Keywords	Deployment Environment, clinical pilots
Editor Address data	Name: Jérôme Lhaut Partner: Institut Jules Bordet Address: Boulevard de Waterloo 121 Phone: +32 2 541 31 11 Fax: +32 2 541 35 06 E-mail: jerome.lhaut@bordet.be
Delivery date	

0.4 Distribution list

Date	Issue	E-mailer
18/12/2012		fp7-integrate@listas.fi.upm.es
17/01/2013		fp7-integrate@listas.fi.upm.es
18/02/2013		fp7-integrate@listas.fi.upm.es
01/03/2013		fp7-integrate@listas.fi.upm.es

Table of Contents

1 Introduction

The main purpose of this deliverable is to report on the preparation of the deployment environment by defining verification and validation procedures that are required to test the INTEGRATE platform under legal and security requirements. The Integrate tools are planned to be evaluated and validated at the *Institut Jules Bordet* and at additional external collaborative clinical pilot sites.
The Extraction-Transform-Load (ETL) guidelines describe the process required to add new data sources to the INTEGRATE platform and to the Common Data Model (CDM) for semantic interoperability.
In this document we define the quality procedures for the validation of the scenarios that are part of the demonstrators. Measurable elements are useful to prove the valuables uses of the platform. Installation qualification (IQ), operational qualification (OQ) and performance qualification (PQ) define validation activities.

2 Summary

The objective of this task is to prepare the technical and procedural environment – in compliance with the defined legal and security framework of the project – for the installation of INTEGRATE technologies and tools for their extensive evaluation and validation. It is also its responsibility to design, oversee and execute all activities, including training, for preparing the clinical pilots for their validation activities.

3 Context

3.1 Terminology

This first section aims to clarify the terminology and in particular the concepts of validation and verification. Many software engineering journal articles and textbooks use these terms interchangeable as if it is a single concept, with no distinction between both terms. However, it is important to understand the difference between these two distinct but complementary activities.

3.1.1 Verification

The General Principles of Software Validation [1] defines verification as "*Software verification provides objective evidence that the design outputs of a particular phase of the software development life cycle meet all of the specified requirements for that phase. It looks for consistency, completeness, and correctness of the software and its supporting documentation, as it is being developed, and provides support for a subsequent conclusion that software is validated. Software testing is one of many verification activities intended to confirm that software development output meets its input requirements. Other verification activities include various static and dynamic analyses, code and document inspections, walkthroughs, and other techniques.*"

Generally, the verification consists in evaluating whether a product, service, or system complies with a regulation, requirement, specification, or imposed condition or not. In the domain of Computer Science, it is the process of determining that a model implementation accurately represents conceptual description of the model.

3.1.2 Validation

The validation is the guarantee that a product, service, or system meets the defined needs and other identified stakeholders. Regarding validation, FDA defines it as follow: "*Confirmation by examination and provision of objective evidence that software specifications conform to user needs and intended uses, and that the particular requirements implemented through software can be consistently fulfilled.*"

The software validation activities may occur both during and at the end of the software development life cycle to ensure that all requirements have been fulfilled. It is required to assure the quality of the developed software system and can increase the usability and reliability and decrease failure rates. This process is highly dependent of software testing, analyses, and verification tasks.

In practice, validation process is executed through the use of procedures defining specific operations that must be taken to complete individual validation activities, tasks, and working items. In our context, these procedures will be formalized in the form of guidelines and protocols describing the IQ *(installation qualification)*, the OQ *(operational qualification)*, and the PQ *(performance qualification)*.

3.2 Legal compliance

Before medical data are transferred to the Integrate data warehouse and therefore be embedded in the Integrate Data Protection Framework, in order to process such medical data from patients of the *Institut Jules Bordet*, some legal and internal requirements are to be fulfilled.

Requirements of the Belgian Law on Data Protection (Law of 8 December 1992)

In principle, the processing of medical data is prohibited (article 7§1), except if such processing satisfies one of several conditions (article 7§2) as foreseen by the Belgian Law.

- One exception is the **consent** of the data subject/patient (article 7§2, a). For living patients, the IJB has developed a broad informed consent (the « Patient Information and consent to use biological and medical data for scientific research ») which has been given positive opinion of the *Institut Jules Bordet* Ethics Committee on July 7, 2011. This consent is explicit but not specific. However, even if not specific, the processing is compatible with the finality of an integrated mono-specialized cancer public institution where a patient can reasonably foresee that such processing could occur. The Belgian law remaining silent about the processing of personal data from deceased patients, this issue will be addressed in next section.
- Another exception to this prohibition to process medical data is **scientific research** (article 7§2, k). If the processing is necessary to scientific research, data may be processed (upon conditions). Scientific research not being defined by the Belgian Law, the Belgian Data Protection Authority has interpreted it in a broad way as any research based on an objective method (objective observations and measurements and statistical analysis) that has a scientific purpose. The INTEGRATE project may satisfies this condition as well.

Requirements of the *Institut Jules Bordet*

In addition to these legal requirements, the *Institut Jules Bordet* has defined internal policies that researchers must comply with. These are not legally binding.

- Each employee of the IJB processing personal data from patients shall be bound by an obligation of **confidentiality**.
- Each processing of personal data must be assessed and given a positive opinion of the IJB **Ethics Committee**. In its review process, the Ethics Committee shall take into consideration the interests at stake balancing the scientific and ethical aspects of the project. With regard to the secondary use of personal data collected within the frame of the TOP Study, the Ethics Committee has been given its positive opinion on July 7, 2011 on the processing itself, the information to the living patients and the processing of personal data from deceased patients.
- The researchers shall process **coded/pseudomymised** personal data and shall take all the necessary and appropriate IT and security measures to

protect such personal data. For the pilot phase, the personal data will be pseudonymised through an encrypting algorithm and the data transfer secured through a VPN encrypting.

- Each transfer of personal data shall be covered by a **Data Transfer Agreement** between the provider and the recipient.

3.3 Security requirements

3.3.1 Hosting Security Requirements

Everyone involved in the pilot setup should be aware that the pilot environment will deal with real patient data. Although it is challenging in a research project, the pilot environment should be managed as much as possible as a production environment.

The following security requirements give an example of common security guidelines for production environments. The mentioned requirements are valid for all servers involved in the INTEGRATE platform, regardless whether they are centrally or locally hosted:

- Physical access to servers must be restricted to authorized personnel only.
- Server administration must be restricted to authorized personnel only.
- Servers should be based on a minimal install.
- Servers and software should be configured according to the least privilege principle.
- Servers must be firewalled according to the least privilege principle.
- OS and core software packages must be kept up to date with respect to security patches.
- Servers should run anti-malware software
- Servers dealing with personal data must implement the security measures deemed required according to the local policy for dealing with personal data which might be organisation dependent (e.g. requiring encryption of data at rest).

In the pilot environment one should not rely on the INTEGRATE security framework for auditing, but rather ensure that “classical” logging is available on all involved servers and applications. Pilot sites are responsible for appropriately shielding the pilot environment from their operational environment. Contact details of administrators of all involved sites should be readily available to the INTEGRATE partners. Users or administrators who detect suspicious behavior, a breach or a bug relating to security should immediately inform the administrators of the involved sites (and the person(s) responsible for the software if a bug was detected). If there is a real risk for data exposure, the pilot must immediately be suspended until a fix is provided.

3.3.2 INTEGRATE Security Framework

Within the INTEGRATE project a security framework specification is part of the architecture. As much as possible, pilot services should rely on the INTEGRATE security functionality. However, one needs to be aware that all software is produced in a research setting, including the security framework. The security software might still contain flaws, or services might have incorrectly integrated it.

Authentication functionality and a standard set of security attributes for access control will be available for the pilot environment. This should allow conducting the pilot securely in a closed community. However, if for a reason, a service cannot be considered secure (e.g. because of incomplete implementation, doubt about the robustness of the implementation), it can only be used in a sandboxed environment during the pilot (e.g. closed VPN).

3.3.2.1 Identity and Access Management

User management for the INTEGRATE pilot platform service is centrally organised. In order to make use of the shared services, partner organisations will need to assign an administrator for their domain or delegate the central platform administrator to do so on their behalf.

Partner organisations are responsible for correctly managing the user information for their users. This includes standard user management tasks such as:

- Ensuring that accounts are only assigned to authorised persons (physical verification).
- Ensuring that where applicable the correct security attributes (e.g. for access control) are set for a person.
- Immediately modifying security attributes or registration information if a person's relation to a partner organisation changes (e.g. leaving the organisation = disabling the account).

INTEGRATE services hosted and managed at the clinical sites that connect to the central platform for information, for example the local EHR data warehouses in the screening scenarios, should be registered to the central security infrastructure. The administrators of these services are responsible for managing identity information of these services ("non-human security principals" in the identity manager).

3.3.2.2 Confidentiality

All communication with INTEGRATE services (and between services) should be authenticated and encrypted as defined by the INTEGRATE architecture. More detail can be found in deliverable “D2.5/4.3 - Integration Guidelines, Initial specification of privacy enhancing services” which describes integration of services with the security framework. In summary:

- Browser authentication must be based on the SAML authentication protocol, supported by the INTEGRATE Identity provider services.
- Confidential (i.e. non-public pages) browser communication should be encrypted through SSL (https).
- Web Service communication must be encrypted at the transport or message level. The mechanism used should be clearly advertised in the WSDL security policy (WS-Security Policy).
- Service requestor authentication during web-service communication is to be done according to the WS-Security SAML Token profile. The necessary SAML credentials can be obtained through the INTEGRATE Identity provider services. Service providers can be authenticated based on standard SSL server side authentication for transport authentication.

For the Pilot environment service providers are responsible for their own access control implementation based on the provided security credentials.

4 Pilot sites

4.1 Institut Jules Bordet

The *Institut Jules Bordet* (IJB) is an autonomous comprehensive cancer centre devoted entirely to the fight against cancer. The strength of IJB is built upon the integration of the following three missions: excellence of care, innovative research and high level of education, and a multidisciplinary approach to the treatment of each individual patient. Accelerating the process through which laboratory findings are implemented in clinical care is IJB's principal goal for the future, with the ultimate aim being to cure cancer.

The original aspect of oncological practice at the *Institut Jules Bordet* is based upon the truly multidisciplinary therapeutic approach, reinforced by a single patient file that is used by all physicians regardless of their specialty. This file is present at all consultations and hospitalizations and collects all medical information on the patient. After establishment of the diagnosis, decisions about treatment are taken by common agreement, by a collective reflexion of surgeons, radiologists and internists, based upon the most modern therapeutic techniques.

The development of today's oncology would not have been possible but for an intense laboratory and clinical research. The "Laboratory for Research and Clinical Investigation" is pursuing fundamental research, especially in the field of haematology, mammary oncology, on melanoma and pulmonary cancer *etc.*, activities that have led to hundreds of scientific contributions recognized at an international level. A fundamental aspect of the *Institut Jules Bordet* is the close integration of research and practical oncology; as a consequence clinical research is permanently present in the treatment the physicians apply. A number of these clinical research programs are carried out in cooperation with national and international centres, especially in the framework of the EORTC (European Organization for Research and Treatment of Cancer) that was founded at the *Institut Jules Bordet* in 1964.

Academically and in the framework of the missions entrusted to the *Institut Jules Bordet* by the *Université Libre de Bruxelles*, the Institute has gained a reputation for the quality of its teaching of graduate and post-graduate courses. Every year the Institute forms and trains several oncologists who later practise their specialities in hospitals of Belgium as well as young researchers who will return to their respective hospitals to make use of their newly acquired expertise.

4.2 The extension of the INTEGRATE consortium with the inclusion of new clinical groups and validation sites

The main development of the INTEGRATE data sharing environment and of the tools to support clinical research was carried out together with BIG and IJB at their sites, based on the clinical requirements provided by the clinical users in these organizations. Our vision was to build solutions based on widely adopted healthcare standards and terminologies that can be deployed and will address the needs of a clinical community much wider than the INTEGRATE consortium. To support future adoption and keep a wide view on the needs of clinical research in oncology we have

established collaborations with several research organizations that share our vision to further enhance data sharing to support research. Next to that, we took an iterative approach of development and prepared an extensive validation of our solutions.

In the validation phase and the final development iteration that starts in the third year of the project we aim to involve several new clinical organizations in the project. To this end, from the start of the project (in the proposal) we have reserved a budget for funding new clinical sites that would take an active role in the validation phase of the project. The goal is to evaluate the scalability of the developed solutions and already during the running of the project make sure that INTEGRATE is able to suit the needs of a large community of users which is not limited to the initial project partners and external collaborators. Therefore, we aim to involve additional clinical organizations throughout Europe and beyond in the pilots of the project.

Such an approach has many benefits. We believe that we address global issues that are relevant for a wide community and the interest of additional top clinical centres to join our team and contribute to the project emphasizes the importance of the project's objectives. The inclusion of requirements from a larger user group with expert knowledge will foster the generalizability of the approach and enable us to indeed address global needs; we will be able to extend, adapt and improve our solutions during the duration of the project. Access to additional data and clinical systems will be used to evaluate the scalability of the INTEGRATE tools. Finally, this larger clinical community will also support future adoption and sustainability, and enable us to demonstrate early enough that the results are valuable for a wide community of clinical users beyond the initial consortium.

During the second year of the project we have reached out to several top research organizations and presented the project objectives and results to different audiences in several clinical and medical informatics events. We have selected two top research groups with extensive expertise in breast cancer research and invited them to become part of our team. One of these organizations is European: the German Breast Group, the second is the Australia and New Zealand Breast Cancer Trials Group (ANZBCTG). Selecting a research organization outside the EU has the role to strengthen International cooperation and to demonstrate the validity of our solutions beyond the European context. It also enables us to benefit of the expertise of this very active research organization.

Both new research organization will participate in the validation of the INTEGRATE environment, provide clinical requirements, and contribute data from both care (EHR data) and research (clinical trial data). To provide to the consortium care data and information on care systems deployed they will involve a hospital in their national network.

4.2.1 The German Breast Group

The German Breast Group (GBG), a leading cooperative study group in the field of breast cancer in Germany, provides the comprehensive management of clinical trials in all major therapeutic categories: prevention, neoadjuvant, adjuvant, and palliative.

The vision of the GBG is best described as healing by innovation, competence and partnership, from the protocol design and feasibility assessments to the final study report.

Through project management in combination with the expert data management and statistical analyses, the GBG delivers consistent high quality results in order to improve treatment therapies of cancer patients and their quality of life.

The main focus of the GBG is on the investigator initiated trials (IIT), clinical studies based on the work of doctors conducting research and focused on the optimisation of therapy (TOP-optimal use trials) and the overall improvement of its quality, unlike in the case of industrial studies which are affected by typical approval and marketing aspects.

The GBG currently manages over 40 clinical trials. All services provided by GBG are to the highest standard of The International Conference on Harmonisation of Good Clinical Practice (ICH-GCP1998) and if necessary regulatory requirements.

The GBG offers a comprehensive range of services, including:

- Idea and Conception of Study Design
- Clinical Project Management
- Clinical Monitoring
- Data Management
- Biometric and Statistics
- External Documentation
- Translational Research
- Biobanking
- Pathological Central Laboratory
- Continuous Medical Education
- Medical Writing
- Sponsorship
- Quality Control

In the year 2012, the GBG had 606 participating sites. Five sub-boards were active during 2012 in the fields of neoadjuvant, adjuvant, palliative, and operative therapies as well as in the field of translational research. Members of the sub-boards are all well-known professionals, experienced in treating breast cancer patients and active in the field of breast cancer research and clinical studies.

In 2012, the Translational Research Board successfully completed several projects, developed new ideas for translational research projects and evaluated several of those proposals submitted by third parties. Following up on the previous year, the GBG further increased the amount of samples in the main biobank facilities in Berlin and Heidelberg as well as in the cooperating laboratories in Hamburg and Erlangen.

In almost 30 clinical trials GBG has been collecting:

- formalin-fixed paraffin-embedded (FFPE) tissue,
- fresh frozen tumour tissue,
- RNA later samples,
- full blood samples for SNP analyses,
- circulating tumour cell samples,
- serum samples,
- plasma samples.

The above results and expertise advertise the GBG as a strong partner for the INTEGRATE project.

4.2.1.1 Frankfurt University Hospital

The GBG will participate to the INTEGRATE project together with the Frankfurt University Hospital (http://www.klinik.uni-frankfurt.de/). The central tasks of the hospital are to provide teaching, research and patient medical care at the highest possible levels from both a national and international perspective. Their guiding principle is to provide the best possible medical care for the patients as a result of greater knowledge. Teaching, research and patient medical care are of equal importance to the hospital and are closely related to one another. Therefore, the hospital sees high relevance in enhancing data sharing, improving efficiency of clinical research and closing the gap between research and care, which are key objectives of INTEGRATE.

The hospital has deployed the ORBIS system from AGFA Healthcare as their Hospital Information System. The data that will be provided to the INTEGRATE consortium comprises anonymised patient health records, namely baseline characteristics of patients, and clinical study records.

4.2.2 Australia and New Zealand Breast Cancer Trials Group (ANZBCTG)

The ANZBCTG has as mission to eradicate all suffering from breast cancer through the highest quality clinical trials research. To achieve this goal their vision is to be a global and regional leader in research collaboration.

The group is the largest independent, oncology clinical trials research group in Australia and New Zealand and has conducted clinical trials for the treatment, prevention and cure of breast cancer for over 30 years. The research program involves multicentre national and international clinical trials and brings together over 500 researchers in 78 institutions throughout Australia and New Zealand. This collaboration ensures that knowledge is shared, resources are pooled and progress is faster. The ANZBCTG pursues a policy in its clinical trials program of evaluating treatment efficacy (is there scientific evidence that a treatment works); quality of life (how well do patients tolerate a new treatment); and cost-effectiveness (how affordable is the new treatment for the community). All research conducted by the ANZBCTG is carried out to the highest ethical and regulatory standards.

The ANZBCTG is one of the most successful, respected and longest established breast cancer research groups in the world. The research program plays a pivotal role in influencing breast cancer practice globally contributing to better outcomes for thousands of women in Australia and New Zealand, and potentially millions more throughout the world.

The ANZBCTG has collaborated with the Breast International Group in the past with excellent results. They also believe in the role of ICT to support efficient research and improve secondary use of data, and strongly support enhanced data sharing in clinical research. The ANZBCTG can support the dissemination and adoption of INTEGRATE results beyond the EU context and enable INTEGRATE to provide solutions with global impact.

5 Validation process

5.1 Definition ETL guidelines

This section describes the process required to incorporate new data sources to the INTEGRATE platform and the Common Data Model (CDM), where syntactic and semantic issues need to be addressed to achieve semantic interoperability. In the current development, the following standards have been adopted: HL7 RIM as CDM, and SNOMED CT as foundation of the core dataset extended with a set of specific vocabularies for laboratory test (LOINC), for adverse events (MedDRA) and for molecular information (HGNC). The Extract, Transform and Load (ETL) process requires, in this environment, syntactic and semantic validation steps described below.

The first step required for data providers is to supply the required documentation to understand the data:

1. Documentation describing the data: A description of the data is required in order to understand and store each piece of information in the correct place at the CDM. Mainly the information required is the description of data fields and values
2. Annotations of data on domain vocabularies (ICD, NCI Thesaurus, SNOMED, LOINC, MedDRA, HGNC, etc.)
3. Comprehensive subset of data (if not complete): At least an example of the data is needed to test and check the format
4. A set of usual queries: What kind of information users need to retrieve from the CDM. Examples can be provided in free text (e.g. “Retrieve breast cancer patients with an age less than 25 at the time of diagnosis”).

The ETL process will start with HL7 v3 messages provided by the new data source (there exist open source tool such as Mirth Connect to perform this task). After that, the file is validated against a schema XSD, detecting syntactic errors and some semantic inconsistencies.

The process aims to incorporate normalized data sources, but in practice, data providers need to solve some issues where interoperability standards are not yet complete. The proposed process provides an interface for the data provider in order to introduce the data through the processes of normalization and binding of terms that follow the standards and also store the original data allowing the user to query for the data as they originally stored it. This process is shown in **Erreur ! Source du renvoi introuvable.**.

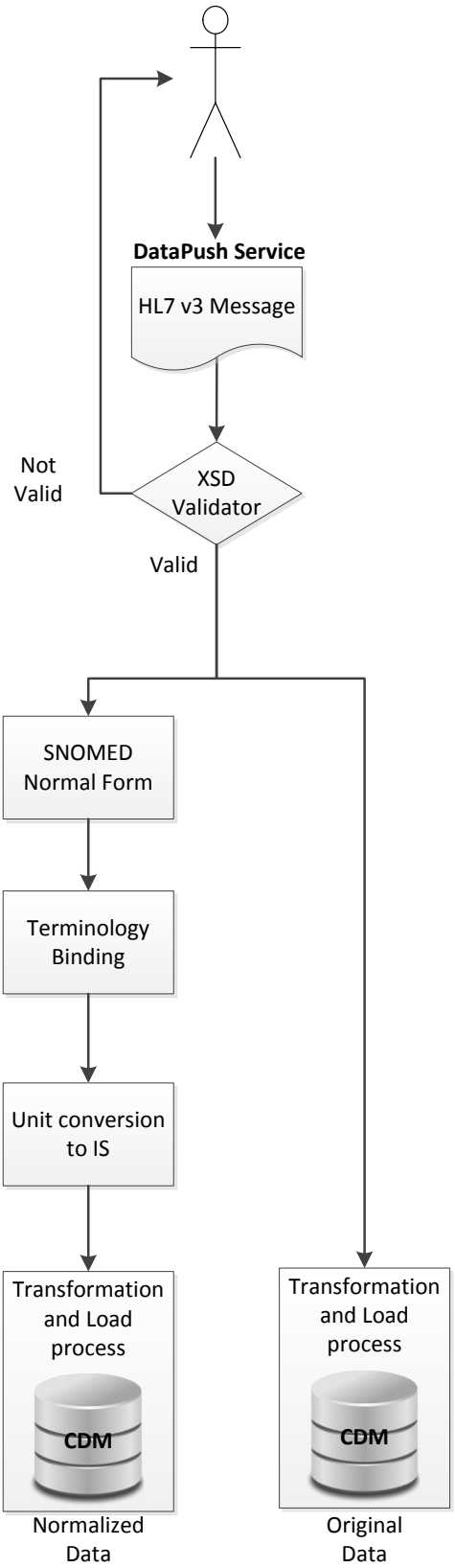

Figure 1 ETL guidelines to include new data sources to the INTEGRATE Common Data Model

The first phase of the process validates the source XML against a XSD template. If the XML is valid, the process continues to the next step, if not, the process returns an error message to the data provider. After the validation step, the process flow is divided in two different branches. One of them will store the original data provided in the Data Warehouse without changes. The other branch will perform a series of transformations and calculations to store the data in the Data Warehouse. The first transformations are performed by the SNOMED normalization process which will transform the incoming SNOMED concepts into their normal form. Then, the Terminology Binding process which will assign a class from the CDM where the normalized concepts can be stored (For example, “leg” should be stored as a “targetSite” and not as a “Procedure”). The Terminology Binding can receive concepts from different vocabularies (SNOMED-CT, LOINC, MedDRA and HGNC) so it needs information about each terminology in order to select the correct class of the HL7 RIM. Finally, this branch will also perform a process of conversion of the units to the International System (IS) allowing the information to be stored in the same measuring units and allowing comparisons of those values.

In both branches, a final step is performed to load the data, the transformation and load part of the ETL process. These will perform some minimum required transformations in the data and load that data into the Data Warehouse.

The purpose of having the branch to store the original data is to allow the data provider to keep traceability and link to the legacy system, avoiding lost information. Such branch makes the data provider responsible of the data semantic interoperability issues that may be introduced (e.g. if the concept “Trocar biopsy” is introduced without normalization it would be impossible to get the procedures performed with a trocar device). This is better than not being able to introduce the data at all although some semantic capabilities will be lost. Despite this, the presence of the “normalization” branch allows to perform an analysis of the data in order to obtain the required semantic interoperability. Although the lack of completely fully defined interoperability standards for such multi-scale and changing domain (research on cancer) may prevent complete interoperability in real cases, the proposed ETL guideline exploit the semantic capabilities available nowadays.

5.2 Assessment/ validation of the benefits of the platform

In this section, each partner responsible for a demonstrator has to supply measurable elements in order to prove the benefits of the platform. It's also required to propose validation activities – IQ, OQ and PQ.

5.2.1 Cohort selection

5.2.1.1 Measurable elements

The cohort selection application aims at providing researchers a tool for easily scanning and filtering large patient datasets based on medical characteristics (cf. deliverable D1.5 "Consolidation of the User Needs, Use Case Development and Requirements Analysis (final)", section 3.7). At the time of writing, the cohort selection application is still in an early stage of development. The work is concentrating on tackling a number of difficult scientific and technical questions. Current activities include defining the query language, developing the query engine and further extending the CDM/CIM access layer.

With respect to the cohort selection application, the following measurable elements can be considered relevant:

- Correctness of the query results
  - Expected vs. obtained filter result
- Query execution speed
  - Performance dependence on data volume

Next to those easily measurable elements, user-friendliness is one of the major determining factors for success of the cohort selection tool. Especially the way to create "filters" (queries which define a cohort) is determining for the overall user experience. At this point in time however, the aspect of query authoring has not yet been addressed in the project. Queries currently need to be written using a Domain Specific Language ("Snaggletooth Query Language"), which is convenient for expert usage, but not for the average end-user.

5.2.1.2 Extending validation activities

Installation qualification (IQ)

The application is still in an early design phase, installation qualification (IQ) is preliminary.

The cohort application currently under development consists of a query execution engine (java application) which is instantiated as a web service ("Cohort Engine Service"). A minimalistic front-end for functional testing purposes is available as a separate web-application. The final application will include a stand-alone front-end (same approach as with the patient screening application) with extensive functionality for managing cohorts and support for authoring queries.

The final cohort selection application will depend on:

- The INTEGRATE platform security infrastructure

- The Query Builder Service (cf. deliverable D2.6 “System Architecture Refinement, Security Framework and Implementation Status”)
- The Core Dataset Service (cf. deliverable D2.6 “System Architecture Refinement, Security Framework and Implementation Status”)
- A clinical datawarehouse accessible through the INTEGRATE CDM/CIM access layer (source for cohort selection)

The final installation qualification for the central Cohort Engine Service will roughly be:

Step	Procedure	Expected results
1.1	Deploy & configure OS	An operational and ready to use minimal install OS
1.2	Install Java JDK (latest edition)	Required Java Runtime environment and libraries for running the application to be deployed properly
1.3	Install Groovy runtime (latest edition)	Required Groovy Runtime environment and libraries for running the application to be deployed properly
1.4	Deploy Tomcat (latest version)	Availability of a Tomcat application server
1.5	Install Cohort Engine Service code	Running application
1.6	Add screening service configuration to the security infrastructure	Inclusion of the screening service into the operational platform (when the service is not registered in the security environment, it cannot address or be addressed by other services)
1.6	Verify access to the Query Builder Service and Core Dataset Service	Connectivity to the CDM/CIM support services
1.7	Verify access to the data source (CDM/CIM enabled data warehouse)	Connectivity to the data warehouse

Operational qualification (OQ)

At this point in time it is too early to detail the operational qualification (OQ) as the main workflow of the application is not yet finalised. An indication of workflow is given in the technical use cases “UC.TQ.*” of deliverable D1.5 “Consolidation of the User Needs, Use Case Development and Requirements Analysis (final)”.

Performance qualification (PQ)

Step	Procedure	Expected results
1.1	Execution of different medical	• Regardless of stress, the

	queries on varying sizes of datasets.	application should not crash • Performance measure useful for tuning (e.g. hardware scaling)
1.2	Simultaneous use of the cohort selection application by multiple users	• Performance degradation will occur, however the application should not crash • Performance measure useful for tuning (e.g. hardware scaling)

5.2.2 Patient screening

The patient screening application is the main application described in the molecular testing use cases explained in deliverable D1.5 “Consolidation of the User Needs, Use Case Development and Requirements Analysis (final)”.

5.2.2.1 Measurable elements

The patient screening application aims to (partially) automate an otherwise (tedious) manual process. Finding a trial in which a patient can be enrolled requires an investigator to examine a large set of eligibility criteria for that patient. The application assists in automating lookup of medical data and comparison (patient vs. criteria).

Validation can be done based on the following measurable elements:

- Correctness of automated criteria evaluation.
- The overall duration of screening process using the screening application versus the classical process (without tool).
    - For this, the patient screening is ideally put in direct competition with the manual process in a real pilot.
- Increase in recruitment rate when using the screening application (the tool is expected to give a better coverage of all running trials than a human could).
- More optimal recruitment (i.e. best fitting trial taking into account the patient benefit and overall research requirements).
    - The latter two measurable elements can only be properly evaluated after a considerable period of real usage.

Finally, as the tool is trying to improve an existing process (by assisting a human), ease of use is a primary success factor of the application. This can be measured through user surveys.

5.2.2.2 Extending validation activities

Installation qualification (IQ)

The patient screening application consists of two Java applications which need to be deployed on the platform as services (“Patient Screening Service” and “Criterion Matcher Service”) and a .NET front-end application which needs to be installed on the computer of the end-user. The .NET front-end communicates with the “Patient

Screening Service", which provides the main application flow and relies on the "Criterion Matcher Service" for evaluating criteria based on patient data.

The patient screening application further depends on the availability of an operational platform security infrastructure and on the following deployed services:

- Trial Management Service
- The platform screening datawarehouse (accessible through the INTEGRATE CDM/CIM access layer)
- One or more EHR datawarehouse (accessible through the INTEGRATE CDM/CIM access layer), depending on the end-user affiliation.

Platform Installation (screening application)

Step	Procedure	Expected results
2.1	Deploy & configure OS	An operational and ready to use minimal install OS
2.2	Install Java JDK (latest edition)	Required Java Runtime environment and libraries for running the application to be deployed properly
2.3	Deploy Tomcat (latest version)	Availability of a Tomcat application server
2.4	Install "Patient Screening Service" and "Criterion Matcher Service" code	Running application
2.5	Add service configurations to the security infrastructure	Inclusion of the screening and matcher service into the operational platform (when the services are not registered in the security environment, they cannot address or be addressed by other services)
2.6	Verify access to the platform Trial Management Service	Connectivity to the platform Trial Management Service
2.7	Verify access to the platform screening datawarehouse	Connectivity to the necessary data warehouses

Client Installation (on an investigator's computer)

Step	Procedure	Expected results
2.1	Run the screening application installer	Working patient screening application on a local computer
2.2	When necessary (this configuration is typically included in the installer) configure the network location of the platform Trial Management Service, screening datawarehouse and patient screening service	Completed platform configuration for the screening application
2.3	Configure the network location of the local EHR data warehouse	Fully operational screening application

Operational qualification (OQ)

Step	Procedure	Expected results
2.1	An investigator starts the screening application (on his computer) and logs in	• The application authenticates the user with the provided credentials (on the INTEGRATE platform) • After successful login, the end-user is presented with the "patient selection screen" • Investigators can only see patients from their centre
2.2	The investigator selects a patient from the list and subsequently chooses which trials the patient should be screened for	• The screening service is started and uses the data of the Trial Management Service and available warehouses to check in how far the selected patient matches the eligibility criteria • The end-user is presented with the result of the patient screening
2.3	The Investigator selects one of the trials to examine the results of the screening in more detail	• A detailed overview (matching result with the patient data) of the eligibility criteria of the chosen trial is presented
2.4	The investigator overrides one of the automated screening results	• Manual override takes preference upon automatically evaluated criteria. Subsequent "screenings" of this patient will take the manually entered information into account
2.5	One of the supporting services (Trial Management Service or data warehouse) should be turned of	• The screening application should not crash and provide proper error detection
2.6	User rights should be revoked by Administrators	• The screening application should no longer function and indicate "access denied" to end-users

Performance qualification (PQ)

Step	Procedure	Expected results
2.1	Eligibility screening of a patient for a large number of trials with a large set of eligibility criteria	• Performance degradation will occur, however the application should not crash • Performance measure useful for hardware scaling
2.2	Simultaneous use of the screening application by multiple users	• Performance degradation will occur, however the application should not crash • Performance measure useful for hardware scaling

5.2.3 Analysis tools

The INTEGRATE Analysis Platform (IAP) provides a framework with applications for the analytical tools for the statistical analysis of a cohort, and the sharing of predictive models for cancer prognosis and treatment response. To enhance functionality the IAP is coupled with an intuitive clinical data browser that takes as input query results and filters the data according to any of the available parameters (i.e. clinical or genomic) through a friendly user interface (UI). Subsequently, the user can select specific columns over the filtered cohort of patient and run tools by just pressing a button. The results of the analysis, including the filtered patient cohort, are presented in a dynamic report that can be stored together with its metadata information in a history record used for future reference.

5.2.3.1 Measurable elements

Several measurable elements from the point of reducing the computational time, cost and user's effort along with the offering usability of the platform can be defined and mentioned below. From a technical perspective, these are:

- Time for implementing an analysis based on a defined research question as addressed in D1.2. To measure this process, several tests will be performed using different cohorts each time. Half of the cases will be reviewed using separate platforms for
    - implementing, or running a tool if it is available
    - programming source code for creating the dynamically statistical record
    - storing the overall metadata information into a storage place as a record

  Half of the cases will be reviewed using the platform. The steps and time required for the analysis, and the usability of the platform will be measured.
- Aspects such as the time required and the efficacy for retrieving cohorts from a data-warehouse to the platform, and the time and complexity required for accessing and modifying the analysis reports from the metadata storage of the platform.

A series of surveys of satisfaction and ease of use need to be answered by a group of users not related with these complex procedures. The surveys aim to confirm that the platform offers a fully-operational and a user-friendly manner in which even a user with no IT background can performs an analysis. The issues and questions that compound the surveys will be defined according to the goals and desires of the developers. Despite the fact that such measure is not quantitative enough it provides a way of highlighting the pros and cons of the platform needed for further improvement.

5.2.3.2 Extending validation activities

The validation scenarios for the INTEGRATE Analysis Platform have already been reported in D6.3. To extend this activity, establishing documented evidence which provides a high degree of assurance that the platform meets the predetermined specifications in terms of Installation Qualification (IQ), Operational Qualification (OQ), and Performance Qualification (PQ) is required.

Installation qualification (IQ)

Step	Procedure	Expected results
3.1	Deploy & Configure OS (Linux)	An operational and ready to use OS
3.2	Deploy RDBMS (MySQL 5.X)	An operational Database System to use
3.3	Configure MySQL for Liferay support	A compatible with liferay database environment
3.4	Install Java 6 (latest edition) (JDK)	Required Java Libraries to be deployed properly
3.5	Deploy Web Application Server (Apache Tomcat 7) (*)	Tomcat to be deployed properly
3.6	Configure Tomcat for Liferay support (*)	To have a compatible with liferay application server
3.7	Deploy Liferay Portal 6.1 CE GA 2 (*)	
3.8	Deploy required Java libraries	Libraries to be deployed properly
3.9	Deploy and Configure Analysis Platform portlet	Portlet to be deployed properly
3.10	Deploy SSO extension (Custodix)	To have a secure, single registration and authentication point
3.11	Install R [3] software environment	R to be installed properly
3.12	Install R server for TCP/IP connectivity [4]	TCP/IP connectivity installed properly
3.13	Install R Libraries required for implementing the analysis tools and the integration between R and Latex documentation [5].	Libraries installed successfully
3.14	Install Latex editor platform [6]	Latex editor installed successfully

Installation qualification (IQ) involves verifying complete arrival of the system as purchased, with a list of components, instruments, and required specifications to be checked and signed off. Specifically for the analysis platform:

- Hardware
  - **Server**
    - CPU: Quad core INTEL processor or better
    - RAM: 16GB (or more)
    - GPU: Any
    - HDD:
      - A 256GB SSD for the OS and
      - a minimum requirement of 3TB storage space in raid configuration for backup.
  - **Client**
    - Any modern PC with a browser supporting HTML5 is sufficient.

- Software
  - **Server**
    - The server needs to run Linux (the suggested operating system for the server is the latest LTS release of Ubuntu).
  - **Client**
    - Any modern operating system.

- Support Applications
  - **Server**
    - The INTEGRATE Analysis Platform is composed of a computational infrastructure of different environments and languages adopted for implementing the platform's facilities and the connectivity process which allows the interaction between these components. Therefore, several components need to be installed to the server that hosts the platform. The front-end of the platform is based on Liferay portal [2]. The implementation of the statistical tools was performed in R [3] language. To facilitate embedding R functionality in our java-based interface, a client/server concept using TCP/IP protocol [4] was used for the communication between the R system and the end-user allowing the interaction between the platform and the execution framework. The Analytical Tools Platform supports an engine [5] to create dynamically statistical analysis reports by enabling integration of R code and Latex documentation [6]. On-the-fly reporting is therefore generated by combining the programming source code and the corresponding documentation into a single file. An Apache Tomcat server [7] and Oracle Java JDK 6 [8] are also required for running Liferay.
  - **Client**

The latest stable version of Internet Explorer 10+ or Mozilla Firefox 17+ or Google Chrome 23+ or Apple Safari 5.1.7+ is preferable.

Operational qualification (OQ)

Step	Procedure	Expected results
3.1	A researcher logins the portal using the SSO module	• The platform authenticates the user with the provided credentials • The researcher is directed to the main Analysis Platform's page
3.2	The researcher selects to perform either a statistical or a predictive analysis scenario	• The researcher is directed to the related web-page • A data browser and a drop down menu of all the available tools or models are displayed successfully
3.3	The researcher selects, from a pool of available data, a dataset for analysis	• The platform interacts with the central data-warehouse • The dataset is retrieved and downloaded successfully to the platform • The platform's data browser loads and displays the dataset for analysis
3.4	The researcher filters the data according to any of the available parameters and selects specific columns-variables over the filtered data	• The cohort for analysis is obtained from the filtering functionality of the data browser
3.5	The researcher selects a statistical tool or predictive model and presses its execution button	• A TCP/IP connection is established between Liferay and R • R script implements the analysis • Latex documentation is activated within R • R scripts generate the html and pdf version of the report • A new link that links to the report in pdf format, is appeared under the executed tool or model

3.6	The researcher views and or downloads the produced analysis report	• The report is displayed on the screen or is made available for download
3.7	The researcher gets access to all the analysis records by clicking the "History" tab	• The platform's database is activated • An information table is appeared on the screen with meta-data information of each analysis
3.8	The researcher edits an analysis report	• The html version of the analysis report is displayed • Java tools allow editing the analysis report • A new pdf version of the edited report is generated and stored to the analysis record
3.9	The researcher compares different analysis reports by selecting them from the information table	• The html version of all the executed analysis are displayed vertically • Java tools allow editing the displayed reports

Operational qualification (OQ) includes procedures for testing the system in its selected environment. To meet these requirements, the platform's system will be "stressed out" and benchmarking tests will estimate and evaluate the computational effort and time need to execute all the available statistical analysis tools and predictive models. If failover issues occurred, recovery backups will be activated and tested for their efficacy degree to provide a stable and fully operational system.

Performance qualification (PQ)

Step	Procedure	Expected results
3.1	Transfer a large pool of clinic-genomic data from the central data-warehouse to the platform	• Several datasets are transferred with no loss, no delays, and stored successfully to the platform
3.2	The multiple back-end components of the platform incorporate successfully to perform the analysis	• Liferay interacts with the central data-warehouse via web services and ftp connection • Liferay interacts with R via TCP/IP connectivity • R interacts with Latex documentation platform • Liferay interacts with the platform's database

3.3	Simultaneous use of the Analysis Platform from multiple users	• The platform establishes multiple TCP/IP connections with R
3.4	Perform time consuming and computationally demanding analysis scenarios	• Platform corresponds in a consistent and timely manner

Performance qualification (PQ) is documented verification that a method works for a specific system according to its routine usage. Once the platform has been established to the server, several tests will assess the stability of the platform when running continuously under long periods of time. Back-end processes will keep track of any system failure, warning, and the module's uptime, providing detailed reports to the administrator.

5.2.4 Central pathology review

The Central Pathology Review (CPR) module offers a secure mechanism for remote management, viewing, annotation and reporting for slides of pathology images. More than this, it offers a mean of collaboration among reviewers in order to assess eligible patients for trials based on their pathology images.

5.2.4.1 Measurable elements

There can be defined several measurable elements for assessing the benefits of the platform. They are the following.

- Time for completing the pathology image review using the CPR module: To measure this several pathology images will be used. Half of these images will be reviewed using the CPR and half without the system using other means that pathologist use in their daily work. Then the time to complete the review will be measured. Moreover the level of their satisfaction will be measured

- The number of patients enrolled in a trial: since the ultimate target is to increase the efficiency of patient enrolment into clinical trials the increase in patient enrolment using CPR (and the rest of the INTEGRATE platform) will be measured.

5.2.4.2 Extending validation activities

The validation scenarios for CPR have already been reported in D6.2 [9] at section 3.1.6 and in D6.3 [10]. In this document we extend those in terms of Installation Qualification (IQ), Operational Qualification (OQ), and Performance Qualification (PQ).

Installation qualification (IQ)

Step	Procedure	Expected results
4.1	Deploy & Configure OS (Linux)	An operational and ready to use OS
4.2	Deploy RDBMS (MySQL 5.X)	An operational Database System to use
4.3	Configure MySQL for Liferay support	A compatible with liferay database environment
4.4	Install Java 6 (latest edition) (JDK)	Required Java Libraries to be deployed properly
4.5	Deploy Web Application Server (Apache Tomcat 7) (*)	Tomcat to be deployed properly
4.6	Configure Tomcat for Liferay support (*)	To have a compatible with liferay application server
4.7	Deploy Liferay Portal 6.1 CE GA 2 (*)	
4.8	Deploy required Java libraries & frameworks (Hibernate, JSF & Primefaces, Open Layers)	Libraries to be deployed properly
4.9	Deploy and Configure CPR portlet	Portlet to be deployed properly
4.10	Deploy SSO extension (Custodix)	To have a secure, single registration and authentication point
4.11	Deploy and Configure Portlet for Pathology Images Transfer from Data Warehouse (SOAP Service)	Portlet to be deployed properly
4.12	Deploy and Configure Image Tiling (Tile Generator) Service	Service Components to be deployed properly

(*) if liferay bundle is used, steps 4.5 & 4.6 should not be performed since a pre-configured tomcat instance is included and deployed upon liferay installation

The objective of this validation activity is to provide documented evidence that installation and configuration of hardware, operating system, application software and system support applications have been correctly installed per the appropriate procedures. Specifically the CPR requires the following in order to run:

- **Hardware**

Server

- CPU: Quad core INTEL processor or better
- RAM: 16GB (or more)
- GPU: Any
- HDD:
  - A 256GB SSD for the OS
  - Minimum requirement of 3TB storage space in raid configuration for backup.

Client

- Any modern PC with a browser supporting HTML5 is sufficient
- OS: Any modern operating system
- Monitor: Support for Full HD resolution (or better)

- **OS**

Server

- Linux OS (the suggested operating system for the server is the latest LTS release of Ubuntu)

Client

- Any modern operating system

- **Support Applications**

Server

- Oracle Java JDK 6 (latest available version)
- python
- Apache Tomcat 7 server
- MySQL
- Liferay Portal 6.1
- GDAL Library and relevant tools (GDAL: Geospatial Data Abstraction Library)
- Imagemagick library and tools
- FORTH's Tile Generator Service
- FORTH's Central Pathology Review Module

Client

- One of the following browsers at their latest stable version
  - Internet Explorer 10+
  - Mozilla Firefox 17+
  - Google Chrome 23+
  - Apple Safari 5.1.7+

Those should be validated in order to ensure that the system is installed correctly. Several real case scenarios of platform use will be executed, in an extensive mode, in order to verify the error free platform response.

Operational qualification (OQ)

Step	Procedure	Expected results
4.1	A moderator logins to the portal using the SSO module	• SSO registration and login module is working • User roles/groups are assigned as expected
4.2	The moderator executes the task for Synchronization of Images (Transfer) with the Data Warehouse	• Relative SOAP service is working. • Images (raw files and relative Meta data) are transferred and stored from Data Warehouse to CPR environment. • Background Tile Generator Service is run and processes images as expected. • Images are ready for use in CPR environment
4.3	The moderator registers new Image Types, Features and Feature Parameters	• All relative information is stored as expected
4.4	The moderator sets up new associations between elements (Image Types, Features and Feature Parameters)	• Associations between the various parts of information are set as expected
4.5	The moderator sets up a new Review Protocol (R.P.)	• Patients are selected and registered to R.P. • Features are selected and registered to R.P. • Reviewers are selected and registered for a specific R.P.
4.6	A Reviewer logins to the portal environment	• SSO registration and login module is working • User roles/groups are assigned as expected
4.7	The Reviewer goes to his Inbox	• Relevant GUI is working as expected
4.8	The Reviewer selects a pending Image for Review (registered to a specific RP)	• Relevant GUI is working as expected
4.9	Relevant View (GUI) is loaded - Image Viewer & Relative Review Form	• Relevant GUI is working as expected
4.10	The Reviewer uses the Annotation Tools (Image Viewer) in order to mark Areas of interest. Annotation Tools	• Relative GUI and Tools are working as expected

	include elements like markers, shapes drawer, lines drawer, text editor, etc.	
4.11	The Reviewer uses the Review From in order to submit his observations/findings regarding specific Feature Parameters (4.3, 4.4)	• Dynamic Review Forms System work as expected
4.12	The Reviewer submits his review to the System	• Review Form Information (relative to specific Review Protocol) is saved • Image Annotated areas and Annotation Information is saved
4.13	A Reviewer tries to resolve a Review Conflict regarding a specific Image/Patient assigned to a particular R.P.	• Relevant GUI works as expected • Collaboration Tools work as expected
4.14	A Reviewer selects an already reviewed Image for re-review (repeat steps 4.9-4.12) and re-submits his observations/findings	• Review Form Information (relative to specific Review Protocol) is saved • Image Annotated areas and Annotation Information is saved

Establishing confidence that process equipment and sub-systems are capable of consistently operating within established limits and tolerances.

So, CPR will be tested for fail-over, how the backup recovery will be executed in that case and finally about its performance and resilience under load. To do that detailed logs will be kept, in which it will be stored the performance of the system (time to complete an operation). The system will be checked against already measured procedures in our laboratories, and if needed will be optimized.

In case of a failure the system will restart and it will automatically initiate all the necessary services which are needed in order to be fully operational. Once again a mechanism included in the platform, will check the logs for any procedures that have been interrupted, and if any found it will provide a detailed report to the administrator.

Performance qualification (PQ)

Step	Procedure	Expected results
4.1	Transfer a large number of Images (SOAP Service)	• Relevant Information is transferred in a timely and consistent way • Relevant information is stored properly in the Platform

4.2	Process a large number of images (Tile Generator Service)	• Tiles are generated as expected for all Images • Tiles are generated in an acceptable time period
4.3	Simultaneous use of CPR platform from multiple users (Reviewers)	• Platform corresponds in a consistent and timely manner
4.4	Perform repetitive & complex scenarios of platform use	• Platform corresponds in a consistent and timely manner

Establishing confidence that the process is effective and reproducible, this phase will test the ability of the module to perform over long periods of time and under extensive use with tolerance deemed acceptable. In order to test this, the system will be constantly running and logs will keep track of the system faults, the warnings and the errors that will be produced. Moreover the uptime of the module will be measured.

5.3 Requirements for the use of local data

The following sections describe the requirements, mainly under a legal point of view, for the use of data from medical records and clinical studies in context of the Integrate platform. Given the diverse nationalities of the involved partners, the details of the contractual party (relative to data protection regulation and security issues) will not be explained.

5.3.1 Informed Consent Form

A patient who is willing to place his/her data on the Integrate platform has to sign an informed consent allowing the processing of his/her personal data within Integrate. This consent form will explain and define the context and limitations in which the data can be examined, analysed and used. The consent of the patient is needed from an ethical point of view. The patient should be able to determine which data referring to him/her will be processed by whom and for which purposes. For additional information about ICF, you can refer to D1.3 [11].

5.3.2 Contractual agreement

5.3.2.1 Integrate Data Protection Board

Each healthcare organization and/or investigator participating in Integrate will have a contractual agreement with the Integrate Data Protection Board (DPB)[1] concerning data protection and security issues. The agreement between the healthcare organizations and Integrate will define the terms and conditions regarding the processing and storage of the patient's data within their own healthcare organization. They will be responsible for the compliance with both data protection regulations and procedures and policies provided by Integrate. Integrate has to make

[1] Central data controller of Integrate responsible for the compliance of Integrate with current data protection legislation.

sure that the access to the data is protected by the security mechanisms defined in the Integrate framework. Taking into account the multitude of IT-infrastructures and different national legislation, the work involved in drawing up those contracts will be of high importance and substantial.

5.3.2.2 Trusted third party

A trusted third party (TTP)[2] has also to be put in place in order to guarantee the *de facto* anonymous environment created for the Integrate platform. Such TTP has to enter into a contractual agreement with the Integrate DPB as well. This contract must contain rules regarding the storage of the links, the access control to the data base and data security issues. One has to remind that TTP provides a software tool that will perform a second pseudonymisation (the first one being done at the hospitals/investigators level) in the stream of data.

5.3.2.3 End-users

Agreements with Integrate end-users (investigators, institutions and pharmaceutical companies) are needed to grant them access and ensure that they agree with the general terms of the Integrate framework. These will be concluded by the Integrate DPB. These contracts are of fundamental importance for Integrate as the provisions they contain must ensure that only *de facto* anonymous data are processed within Integrate, guarantee patients' rights of access to data and ensure transparency and confidentiality at the same time.

5.3.2.4 Other contractual agreements

For a description of other contracts (Consortium Agreement between the Integrate Partners, contract(s) between an Integrate partner and a sponsor and the contract between CDP and the pharmaceutical companies) required to sharing data, you can refer to D1.3 [11].

5.3.3 Codifying, formatting and obfuscating data

So that the data can be handled by the ETL certain format and coding requirements must be met. Medical data must be formatted according to standard HL7v3 and be identifiable through international domain vocabularies (DCI, NCI Thesaurus, SNOMED, LOINC, MedDRA, HGNC, etc.) Precise documentation describing the different fields is also required for the extraction process.

Once the data are properly formatted, they must be obfuscated in order to be submitted to the platform. This process should help ensuring a high level of security and therefore include the use of modern encryption algorithms (AES, Blowfish, Serpent, RC6, etc.) and sufficiently robust (minimum 256 bits) key sizes. The collected data can then be subjected to the platform where the second stage of anonymisation will take place by a trusted third party before being submitted to the ETL.

[2] In this context, trustful custodian for personal data or the codes/keys/links that identify the data subject and which shall ensure the privacy of the data subject.

6 Conclusions

This document have detailed the validation procedures that cover the quality evaluation of tools within the INTEGRATE platform and environments. Three different indicators have been defined to this end, which precisely describe the installation, operational and performance qualification, together with measurable elements that aim to confirm the benefits of the project platform.

7 References

[1] Food and Drug Administration, “General principles of software validation; final guidance for industry al FDA staff,” 2002.Liferay portal (www.liferay.com)

[2] Liferay portal (www.liferay.com)

[3] The R project for Statistical Computing (www.r-project.org)

[4] Rserve, a binary R server (www.rforge.net/Rserve)

[5] The Sweave tool (www.statistik.lmu.de/~leisch/Sweave)

[6] Latex, a document preparation system (www.latex-project.org)

[7] Apache Tomcat (tomcat.apache.org)

[8] Oracle (www.oracle.com)

[9] The INTEGRATE Consortium, Deliverable 6.2: Evaluation and validation procedures for the INTEGRATE environment,

[10] The INTEGRATE Consortium, Deliverable 6.3: Specifications of the evaluation and validation scenarios for the different components, December 2012.

[11] The INTEGRATE Consortium, Deliverable 1.3: INTEGRATE legal, ethical and regulatory requirements, December 2011.