SERVICES ASSOCIATED TO DIGITALISED CONTENTS OF TISSUES IN BIOBANKS ACROSS EUROPE

Validation plan

Deliverable 1.4

Project acronym:	BIOPOOL
Grant Agreement:	296162
Version:	v1.0
Due date:	Month 6
Submission date:	15/03/2012
Dissemination level:	PU
Author:	Roberto Bilbao (BIOEF)

Part of the Seventh Framework Programme
Funded by the EC - DG INFSO

Table of Contents

1 Document History

Version	Status	Date
V0.1	draft	08/03/2013
V1.0	final	15/03/2013

Approval		
	Name	Date
Prepared	All Partners	08/03/2013
Reviewed	Elena Muñoz,	13/03/2013
Reviewed	Oihana Belar, Arantza Bereciartua, Fabianne Gandon	14/03/2013
Authorised	Roberto Bilbao	

Circulation	
Recipient	Date of submission
Project partners	08/02/2013
European Commission	15/03/2013

2 Executive Summary

This is the deliverable D1.4 – Validation plan, which is under the scope of the task T1.4 inside the WP1. The aim of the task and of the document is to define the validation plan of the created system at different level.

Software validation increases the usability and reliability of the device, in will be done through dynamic testing following the document: “General Principles of Software Validation; Final Guidance for Industry and FDA Staff” by pathologists. It will be done into three phases (Installation qualification, Operational qualification and Performance qualification), which will be separated in three different levels (Module level testing, Integration level testing and System level testing) and will be done in two steps approach (firstly, on a limited subset of data pool based on colon carcinoma and secondly in a complete tumor panel).

On the other hand is important to design the validation of the system as a useful tool of end users. In this sense, pathologist of BIOEF and EMC will check the functionalities of the system playing the principle three end users’ roles which are researcher, teacher and students and pharmaceutical companies. During the second year new biobanks will participate in the validation of BIOPOOL system. For this aim it has been defined:

a) The set of images/data to be analyzed in each step,
b) The chronogram of the validation,
c) The types of the probes to be done
d) The parameters to be taken into account to value how good is the search results.

3 Glossary

API: Application programming interface
BMP: Bitmap image file
CSS: Cascading Style Sheets
D: Deliverable
DCG: Discounted Cumulative Gain
DP: Digital pathology
EMC: Erasmus Medical Center
ERR: Expected Reciprocal Rank
FDA: Food and Drug Administration
FT: Functional Test
HTML: Hypertext Mark up Language
INIT: Initialization
JPEG: Joint Photographic Experts Group
LAN: local area network
M10: Month 10
M12: Month 12
M24: Month 24
PHP: PHP Hypertext Pre-processor
PoC: Proof-of-concept
ROI: Region of Interest
VPN: Virtual Private Network
WP: Work package
XML: Extensible Mark up Language

4 General principles and context

Part of this document is based on the general validation principles that the FDA considers to be applicable to the validation of medical device software or the validation of software used to design, develop or manufacture medical devices.

Planning, verification, testing, traceability, configuration management and many other aspect of good software engineering are important activities to support a final conclusion that software is validated. The software validation and verification activities are conducted throughout the entire software life cycle.

Based on previously mentioned guidance, unless specifically exempted in a classification regulation, any medical device software product developed after June 1, 1997, is subject to applicable design control provisions. This requirement includes the completion of current development projects, all new development projects and all changes made to existing medical device software. Indeed, other design controls, such as planning, input, verification and reviews are required for medical device software. Thus, the corresponding documented results will provide additional support for a conclusion that BIOPOOL software is validated.

Furthermore, computer systems that create, modify and maintain electronic records and manage electronic signature are needed to be validated. In this sense, BIOPOOL system must be validated to ensure accuracy, reliability, consistent intended performance and the ability to discern invalid or altered records.

4.1 Key words:

As it has been described in FDA Guidance, is important to be clear the meaning of some key words that are used in this kind of documents:

Requirement: Any need or expectation for a system or for its software that reflects the stated or implied needs of the customer, and may be market-based, contractual, or statutory, as well as an organization's internal requirements. These can be of a different nature: design, functional, implementation, interface, performance, or physical requirements. Mostly are typically derived from the system requirements for those aspects of system functionality that have been allocated to software and are defined, refined, and updated as a development project progresses. Success in accurately and completely documenting software requirements is a crucial factor in successful validation of the resulting software.

Specification: It is defined as “a document that states requirements.” It may refer to or include drawings, patterns, or other relevant documents and usually indicates the means and the criteria whereby conformity with the requirement can be checked. There are many different kinds of written specifications, e.g., system requirements specification, software requirements specification, software design specification, software test specification, software integration specification, etc. All of these documents establish “specified requirements” and are design outputs for which various forms of verification are necessary.

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. Software verification 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.

Software validation is a part of the design validation for a finished device. As described in FDA guidance it considers to be a "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."

In practice, software validation activities may occur both during, as well as at the end of the software development life cycle to ensure that all requirements have been fulfilled. Since software is usually part of a larger hardware system, the validation of software typically includes evidence that all software requirements have been implemented correctly and completely and is traceable to system requirements. A conclusion that software is validated is highly dependent upon comprehensive software testing, inspections, analyses, and other verification tasks performed at each stage of the software development life cycle. Testing of device software functionality in a simulated use environment, and user site testing are typically included as components of an overall design validation program for a software automated device.

5 Documentation of the software

All partners involved in BIOPOOL project have committed as a quality plan that they will follow the states that for any application or module they have documentation associated; otherwise the implementation is not completed.

Documentation and testing are intimately linked to the code implementation.

5.1 Pertimm:

Pertimm solution has to be installed on servers with CentOS or RedHat 5.X, with 64bits architecture and the following libraries are prerequisite:

- curl 7.15
- bzip2-libs 1.03
- pcre 6.6.2
- mysql-server >= 5.0 (only libmysqlclient must be installed)
- libxml2
- libxslt
- sqlite >= 3
- Ruby 1.9.3-p194 (provided by Pertimm)
- Php 5.3.5 (provided by Pertimm)

5.2 eMedica:

- Microsoft Visual Studio 2010 (in development environment)
- Microsoft .Net Framework 4.0 (in production environment)
- Internet Information Server 7
- DotNetNuke 7.0
- mySQL Server >= 5.0
- ImageMagic
- Google API

5.3 Tecnalia:

- Microsoft Visual Studio 2010
- Microsoft SQL Server
- Matlab 2012
- Microsoft Internet Information Server 7.5

6 Regulatory requirements for software validation

As indicated in the document "Guidance for Industry and FDA Staff General Principles of Software Validation", the FDA's analysis of 3140 medical device recalls conducted between 1992 and 1998 reveals that 242 of them (7.7%) are attributable to software failures. Of those software related recalls, 192 (or 79%) were caused by software defects that were introduced when changes were made to the software after its initial production and distribution. Software validation and other related good software engineering practices discussed in this guidance are a principal means of avoiding such defects and resultant recalls.

Unless specifically exempted in a classification regulation, any medical device software product developed after June 1, 1997, is subject to applicable design control provisions. This requirement includes the completion of current development projects, all new development projects, and all changes made to existing medical device software. Indeed, this document mention that other design controls, such as planning, input, verification, and reviews, are required for medical device software. The corresponding documented results from these activities can provide additional support for a conclusion that medical device software is validated.

In addition, computer systems used to create, modify, and maintain electronic records and to manage electronic signatures are also subject to the validation requirements. Such computer systems must be validated to ensure accuracy, reliability, consistent intended performance, and the ability to discern invalid or altered records.

All production and/or quality system software, even if purchased off-the-shelf, should have documented requirements that fully define its intended use, and information against which testing results and other evidence can be compared, to show that the software is validated for its intended use. The use of off-the-shelf software in automated medical devices and in automated manufacturing and quality system operations is increasing. Off-the-shelf software may have many capabilities, only a few of which are needed by the device manufacturer. Device manufacturers are responsible for the adequacy of the software used in their devices, and used to produce devices.

7 Phases

As described in FDA Guidance, for many years it has been attempted to understand and define software validation within the context of process validation terminology. In this sense, software validation is sometimes described in terms of installation qualification, operational qualification and performance qualification.

7.1 Installation qualification (IQ)

Establishing confidence that process equipment and ancillary systems are compliant with appropriate codes and approved design intentions, and that manufacturer's recommendations are suitably considered.

7.2 Operational qualification (OQ)

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

7.3 Performance qualification (PQ)

Establishing confidence that the process is effective and reproducible

8 Levels

In order to provide a thorough and rigorous examination of a software product, development testing has been organized into levels and each level with a two-stage approach: first on a limited subset of the data pool corresponding to a specific pathology and tissue type, and second on a tumor panel.

8.1 Two stage approach

The BIOPOOL system will be developed, tested and validated in a two stage approach (during M12 and M24):

- The first stage will be the Proof-of-Concept (PoC) model of the BIOPOOL system. This PoC is described in detail in deliverable D6.1. The PoC contains only the minimal requirements of the system and the Digital Pathology (DP) image pool will consist of only one tissue type (colon carcinoma) with a limited amount of images. Testing and validation of this PoC will ensure us that the basic structure is in agreement of what we had foreseen.
- The second stage of the project is the further development of the fully operational BIOPOOL system with all functionalities and an expanded set of DP image pools with multiple tissue types available. Naturally this second stage also needs to be tested and validated.

8.1.1 Stage 1:

The PoC of BIOPOOL and the colon DP image pool are expected to be available for testing and validation in month 10 of the project. A report on this will be provided as deliverable D6.2. The outlines of the PoC and detailed descriptions of all the different components that are part of the PoC, such as the data infrastructure, image and text search engines and the web portal, are described in detail in the according deliverables, as mentioned. The tissue conditions included in the colon DP image pool are limited to carcinoma and adjacent healthy colon. Regarding the functionalities of the BIOPOOL system not all the functionalities have been developed for PoC. For instance, searching can be done either on image morphology or on text searches, but other functionalities as the combined image and text search functionality, together with other extra functionalities, and the increase of different DP image pools, tissue features within each of these pools and thereby the total amount of DP images will be developed in the second stage of the project.

There are different items that need to be validated in the correct order:

- The colon DP images which are used to build the pool must be validated. Each biobank will provide new cases, all with DP images and associated data (based on the already designed tables of histological pattern and anatomic pathological clinical report). All the participant biobanks have agreed to fulfill the minimum characteristics to include the sample/image and associated data in the system. When the DP image pool is validated successfully, only then the next validation step can be done.

- Validation of each of the separate modules of the BIOPOOL system can be done as soon as the development of these modules has been completed. When all these modules are combined to form the PoC, which will be ready in month 10, the testing and subsequent validation on the PoC will begin and will finish in month 12. The results of the validation will be reported in deliverable D6.3. The PoC validation needs to be done under the supervision of the BIOPOOL pathologists.

Therefore the BIOPOOL pathologists from both BIOEF and EMC will perform the testing and validation procedures according to this validation plan. In paragraph 10 of this deliverable the pathologists' requirements validation is described. The key in this PoC validation is to determine if search results will be according to what is expected.

8.1.2 Stage 2:

In the development phase that follows after a successful validation of the PoC, more functionalities will be added to the BIOPOOL system and the amount of different DP image pools together will be increased, together with a more detailed level of pathologies.

As with the stage 1 validation, before the validation of the whole BIOPOOL system starts, first the DP image pools need to be validated in the same way as in stage 1 validation.

During this period other biobanks which are interested on joining BIOPOOL will participate. They will include sets of images and associated data that will be used as a control to validate the system following the steps described above and the next points 8.2, 8.3 and 8.4. Moreover in paragraph 10 of this deliverable the pathologists' requirements validation is described. In this way, the full BIOPOOL system will be validated as a whole, with all functionalities included.

8.2 Module level testing

In this section, the validation plan at module level is described at the development stage of M12. This deliverable containing the validation plan will be revised according to the results over initial tests and objectives of second year during M13 of the project.

This level testing focuses on the early examination of sub-program functionality and ensures that functionality not visible at the system level is examined by testing. Unit testing ensures that quality software units are furnished for integration into the finished software product.

The modules to be tested are Image Search Engine and Text Search engine.

8.2.1 Image Search Engine

Two approaches over the validation of Image Search Engine module are given. On the one hand, first set of tests are related to the functionality itself, i.e., the correct implementation of the functional requirements explained in D1.2; and on the other hand, there must be a second set of tests to evaluate the performance and quality of the results provided by the Image Search Engine.

To this purpose of single module evaluation, Tecnalia has planned to develop an own platform for Image Search Engine testing, before being integrated in the whole BIOPOOL system. The aim of this platform is to validate the quality of the results retrieved by the Image Search Engine, detect possible problems, and tune the indexing (descriptors extraction) and retrieval algorithms to obtain maximum performance of the system. This will serve as testing platform for pathologists in the meantime the whole BIOPOOL architecture is being developed and all single modules are finished and ready for integration.

This platform for Image Search Engine testing is planned to be ready by M10 of the project, this is, June 2013, together with a set of images to carry out the validation tests. Previously intermediate meetings will be held with pathologists to generate the annotated ground truth (by means of which the similarity of images will be characterised and metrics for retrieval purposes will be calculated). This is under the scope of task T4.2. Visual indexing and search module, and the methodology used and results obtained will be gathered in the corresponding deliverable D4.2, scheduled by M12.

8.2.1.1 Functionality validation

The images that will be displayed as results in this set of test will be the original images with full resolution. Restrictions about people executing these tests will be considered (pathologists, authorised people) in order to preserve the diffusion of the images due to legal issues in Spain, and tests will be done in access-controlled rooms. The fact of showing the full resolution image is fully justified by the need of having the feedback of the users about the goodness of the results, whether they fit their expectancies or not. At system level, the images presented back as result of a query won't be full resolution images, and in case the users want to have them, they should contact biobanks that owns those samples.

Just to remind, in this first set of tests (M10 - M12) only one pathology is available, colon carcinoma, therefore it is not needed to incorporate the name of the tissue type in the input query, but in the future it will. Same consideration with mixed search of text and image input. They are not applicable at this stage of the project.

According to the functional requirements exposed in D1.2, the following actions will have to be executed perfectly from the SW platform. The Functionality Tests are indicated next:

FT1. The user can upload its own image stored in local PC into the platform. This will constitute the query image. Maximum size is restricted to 5000 x 5000 pixels

FT2. The user can select a ROI over the uploaded image. This ROI can be rectangle, circle or free line. The part of the image contained inside the selected ROI will constitute the query image.

FT3. The user can launch a search with the query image.

FT4. The user can select certain criteria to refine the search, for example, search by colour, by texture (in definition process at the moment of writing this deliverable).

FT5. The user receives a set of results, placed in a mosaic way, this is, disposed in rows and columns and no less than 9 results (3 rows by 3 columns). The images best fitting the input query are retrieved back, ordered from “most similar” to “less similar”

FT6. The number of results to be displayed can be increased or decreased.

FT7. Every image provided as result will have the possibility (by means of ‘+’ or ’–‘ classification) to be indicated as a positive or bad result, according to user subjective expectations. Relevance feedback is obtained this way even though refine search will not be still possible at this stage of the project.

8.2.1.2 Measurement indicators

8.2.1.2.1 Effectiveness indicators of an image recovery system

In this section we are going to show the effectiveness indicators of the image recovery system. Given a query image the system will return an ordered list of similar images to the user. The effectiveness indicators will evaluate if the generated list contains similar images or not. In addition, the image recovery system needs to be evaluated accordingly to the user satisfaction that the results will generate.

In the scientific and technological community, there are several measurements which are valid for a general Information Retrieval system evaluation. All of them are based on the so called “Ground Truth”, which is the real result of an output. For example, in the case we are evaluating some face recognition system, given an image of a face, the ground truth is the real name of the face. So if the face recognition output is the same as in the ground truth, then the system has a correct output. Given that definition, now it is possible to show some of the most used metrics in the Information Retrieval field. These metrics are *precision* and *recall*.

Precision is the most useful value in applications similar to BIOPOOL. The value of precision shows how many relevant images are in the list of retrieved images. More exactly precision is defined by:

$$precision = \frac{\sum_{i=1}^{N} r_i}{N}$$

Where, *N* is the number of retrieved images and r_i is the binary relevance of the image *i*. In the case of precision, and because of N could be very different, there are several standard cut-off points to give the measurement. Some examples are *precision@10* (with N=10) or *precision@30* (with N=30). As on can see, if we have 5 relevant images in the first 10 elements, we are going to have a *p@10 = 0.5*. If we have more relevant images, we will have a precision value near to 1.

Another less used measurement is the *recall*. This value gives an idea of how many relevant images are in the output list, related to the total number of relevant images. The definition is:

$$recall = \frac{\sum_{i=1}^{N} r_i}{M}$$

Where, *N* is the number of retrieved images and r_i is the binary relevance of the image *i*. *M* is the total number of relevant images in the database. This value can give an idea about if we retrieve all the relevant documents or not. This value is not very useful for BIOPOOL, because for a given list we are not going to know if our output is good enough.

The defined measurements require binary judgements for the recovered images (i.e. relevant, not relevant). Despite this, in the last years there have been proposed several measurements that are able to use graded relevance. One of the most used metric is *DCG* and *nDCG*.

Discounted Cumulative Gain (DCG) gives an idea of the usefulness of a retrieved image which is located in a position in the result list. This usefulness is the *gain* and it is accumulated from the top to the bottom of the list. The definition of DCG is:

$$DCG_N = r_1 + \sum_{i=2}^{N} \frac{r_i}{log_2(i)}$$

Where N is the amount of items in the output list, and r_i is the graded relevance of the image *i*. As it can be seen, this function penalize highly relevant documents (high r_i) that are located lower in the output list (low *i*).

Using DCG, it is not possible to compare several search engines with several queries (because there are different relevance grades). To this end it is necessary to normalize the value for each query, so it is needed to compute the normalized DCG (nDCG):

$$nDCG_N = \frac{DCG_N}{IDCG_N}$$

where IDCG is the ideal/max value of DCG. In this case, the maximum value of nDCG will be 1.0 (when DCG=IDCG). In is worth to mention that this measurement includes some notion of user satisfaction as the usefulness of the judgements, which are dependent on the position of the result.

Other example is the Expected Reciprocal Rank (ERR) [1], where its authors included a definition of how a user acts when the output list of images is present. In this case the following

model of user is supposed: 1) one user asks a query; 2) then the user will navigate through the output list until a relevant document is reached. With this model, the ERR definition is:

$$ERR_N = \sum_{i=1}^{N} \frac{R(r_i)}{i} \prod_{j=1}^{i-1} (1 - R(r_i))$$

where $R(r_i) = \frac{2^{r_i - 1}}{16}$, N is the amount of items in the output list, and r_i is the graded relevance of the image *I* (in Web retrieval this grades are {4,3,2,1,0,-2}).

Regarding the previous measurement definitions, it is clear that a complete Ground Truth is required. This means that these measurements must know the images from the training database that are relevant for a given query image.

BIOPOOL is totally different. In this case, due to the large scale of the problem, it will not be possible to collect ground truth for the entire training database. In this case we cannot use the standard measures (e.g. precision or recall) to validate the BIOPOOL system. In the case incomplete ground truth is available; researchers have proposed new measurements that handle missing judgements differently.

One approximation is to ignore unjudged documents. This is the case of *bpref* measurement that only accepts binary relevance scores [2]. *Bpref* has several improvements, and the most widely used definition is:

$$bpref = 1 - \sum_{i \in R} \frac{(number\ of\ n \epsilon N\ ranked\ higher\ than\ i)}{|\mathrm{R}| \cdot \min\{|R|, |N|\}}$$

Where *R* is the set of relevant judgements in the database (per image), N which is the set of non-relevant judgements in the database (per image), and *n* is a non-relevant document in the output list. In order to compute *bpref*, it steps through <u>all</u> the relevant documents in the output list, and for each iteration it counts the number of non-relevant documents that are ranked higher that the current document.

Besides this definition, authors specified that in the case few relevant document where collected in the output list (i.e one or two), this measurement is very coarse. Because of that, the also defined *bpref-10*, which consists in the same definition, but it considers the top |R|+10 non-relevant documents:

$$bpref - 10 = 1 - \sum_{i \in R} \frac{(number\ of\ n \epsilon N\ ranked\ higher\ than\ i)}{(|\mathrm{R}| + 10) \cdot \min\{(|R| + 10), |N|\}}$$

With the first definition of *bpref* it only takes into account the number of negatives that are located previous to the *R* documents. There have been proposed a different measurement that also makes use of all the negative examples [3]:

$$RankEff = \sum_{i \in R} \frac{(number\ of\ n \epsilon N\ ranked\ higher\ than\ i)}{|\mathrm{R}| \cdot (|J| - |R|)}$$

Where *R* is the set of relevant judgements in the database (per image), *N* which is the set of non-relevant judgements in the database (per image) and *J* is the set of judged documents (*N U R=J*).

As it can be seen, *bpref* only uses binary judgements, but one extension to graded values has been proposed, the *rpref* [4].

$$rpref = \frac{1}{R} \sum_{i \in J} r_i \left(1 - \frac{p_i}{N}\right)$$

Where:

$$R = \sum_{i \in J} r_i$$

$$N = \sum_{i \in J} 1 - r_i$$

$$p_i = \sum_{\substack{k \in J \\ k < i \\ r_k < r_i}} \frac{r_i - r_k}{r_i}$$

R and *N* are normalization values, and p_i is a penalty value that only considers documents that are less relevant but are ranked higher *i* in the output list.

Another approximation to handle the unjudged images is to use information of the process of ground truth gathering. With this information one can modify standard measurements, like *precision*.

One of the most used metrics is the *statMPC*, which definition is:

$$statMPC@k = \frac{1}{k} \sum_{i=1}^{k} \frac{r_i}{\pi_i}$$

Where *k* is the size of the considered output list, r_i is the relevance of the document I and π_i is the inclusion probability.

This definition is very similar to the precision at *k*, but it includes the concept of "inclusion probability". This concept reflects the probability of one document to be selected in the sample of the pool (when considering samples without-replacement).

8.2.1.2.2 Validation measurements

In the previous section we have defined the most widely accepted evaluation measures for an Information Retrieval system. Given these measurements we need to define how many of them we will take into account in order to validate the Image Search Engine, and what the thresholds are for each measure.

In the following table we resume that information.

Table 1: Suggested measurements for BIOPOOL. The references are selected based on the mean results of the TREC Web Track 2012 for statMPC@20, ERR@20 and bpref and rpref of the mean values reported in [4] using the TREC Terabyte data.

Judgement type	Complete judgements	Measurement	Best value	Top performing reference
Binary	No	statMPC	1.0	0.329
	No	Bpref	1.0	0.356
Graded	No	Rpref	1.0	0.698
	Yes	ERR	0.0	0.248

As it could be seen, there are two different set of measurements, depending on the retrieved relevance. At the time this document was written it is not defined the kind of ground truth that it is going to be in the training database. Because of that we have defined two possible scenarios: when only binary judgements are present or when graded judgements are available. In the first scenario, we have selected *statMPC* and *bpref* as effectiveness indicators, and in the second we have selected *rpref* and *ERR*.

In both cases the best value would be obtaining a 1.0 (or a 0.0 for *ERR*). As reference of top performing system we have selected the top performing algorithms that were present in the TREC Web Track 2012 and in the TREC Terabyte track. The average punctuation of those systems is reported in the previous table. We need to highlight that these values are a reference for a top performing algorithms retrieving web scale textual documents. The idea is to be able to compare the values of our system with such large scale algorithms to see if the visual information could ever give as much information as textual.

The results obtained will be included together with the technical description of the visual indexing and search module in the deliverable D4.2. *Report on visual indexing and search module*, to be provided by M12, and the performance evaluation will be reported as well in the deliverable D6.3. *Report on the validation results of proof-of-concept prototype*, scheduled by M12.

8.2.2 *Text Search Engine*

Pertimm has set up a whole testing environment on the basis of continuous integration process. This methodology aims at checking for every change in the source code that it doesn't imply a regression in the delivered executable code.

For Pertimm, this is done in close relation with the adoption of Agile methods in team technical team. The aim has been to obtain a fully automated (no manual action) building chain based on tests tools and a continuous integration system.

The following rule applies to every code implementation: "Each new functionality comes with documentation and test plan".

Everything starts with a version monitoring system (SVN Subversion), a tracking issue system (Redmine, Mantis)

8.2.2.1 For code implemented in C

Architecture has been set up for an automated monitoring and the main components of Pertimm solution have been adapted so that they can be monitored.

This monitoring architecture is composed by 3 parts:

- An XML format enabling to create scenarios, a scenario being a set of instructions describing use case and the criteria defining whether or not the test succeeded.
- A self-made tool (gentlest) converts XML scenarios to executable PERL scripts. Those scripts can be used for an automated execution of the scenarios and the automated generation of log files that can be analysed by other tools (such as Jenkins for continuous integration).
- Monitoring tools include sensors that measure the machines behaviour when running the scenarios

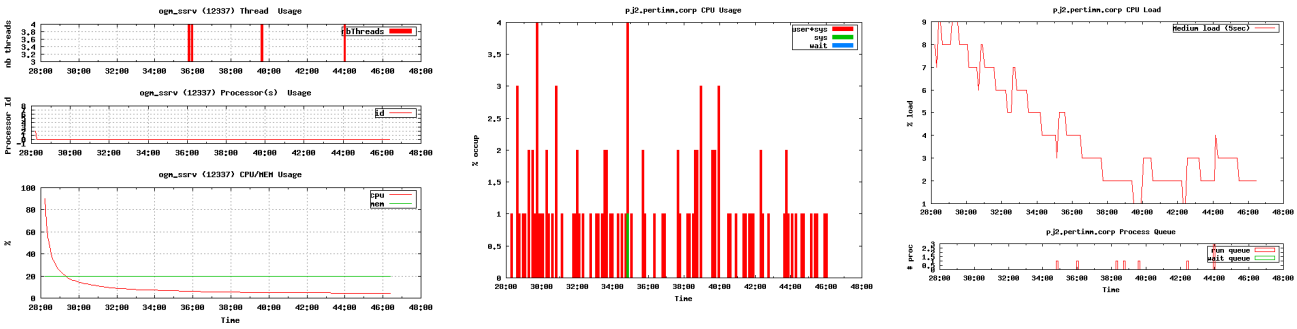


Figure 1: Monitoring

This architecture is connected to a continuous integration server managing the automated monitoring of the tests (scheduled running, alerts, etc.) and the results archiving (Jenkins[1]). It offers a complete up to date overview of the implemented modules:

[1] https://wiki.jenkins-ci.org/display/JENKINS/Meet+Jenkins Jenkins is an application that monitors executions of repeated jobs, such as tests. It provides a continuous integration system. Jenkins makes it easier for developers to integrate changes to a project, and makes it easier for users to obtain a fresh build. The automated, continuous build increases the productivity.

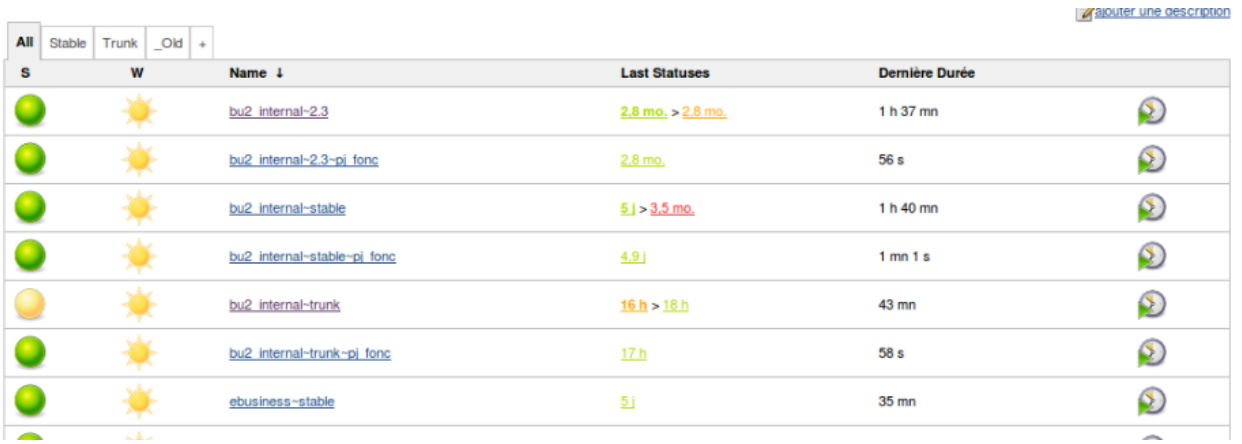


Figure 2: Build results

The following symbols present the state of the four last builds

Symbol	Meaning	Blinking
	Project has never been built before or is not active.	First build is running.
	Last build successfully completed.	Last build successfully completed. And new build is in process.
	Last build successfully completed but is not stable. This is mainly used to present errors on tests.	Last build successfully completed but is not stable. And new build is in process.
	Last build failed.	Last build failed and new build is in process.
	Project health is over 80%.	
	Project health is between 60% and 80%.	
	Project health is between 40% and 60%.	
	Project health is between 20% and 40%.	
	Project health is lower than 20%.	

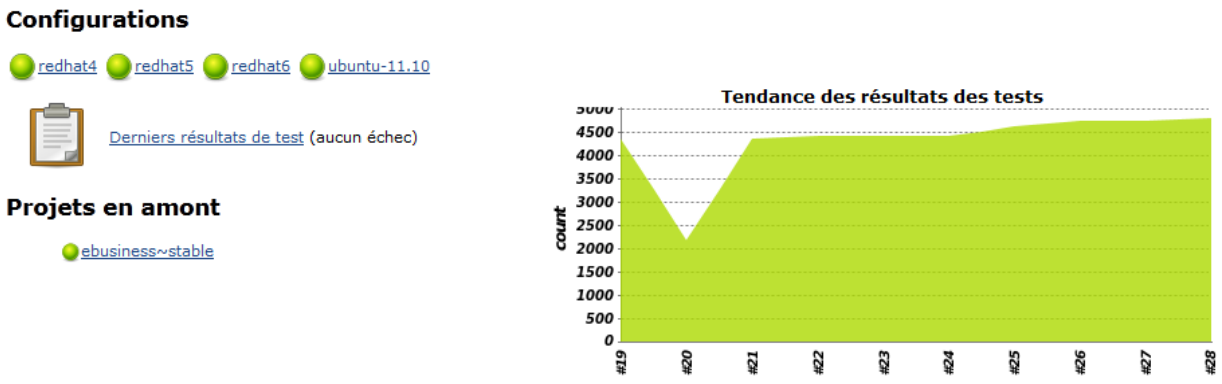


Figure 3: Results

Figure 4: Build partially successful

Figure 5: Current work in process

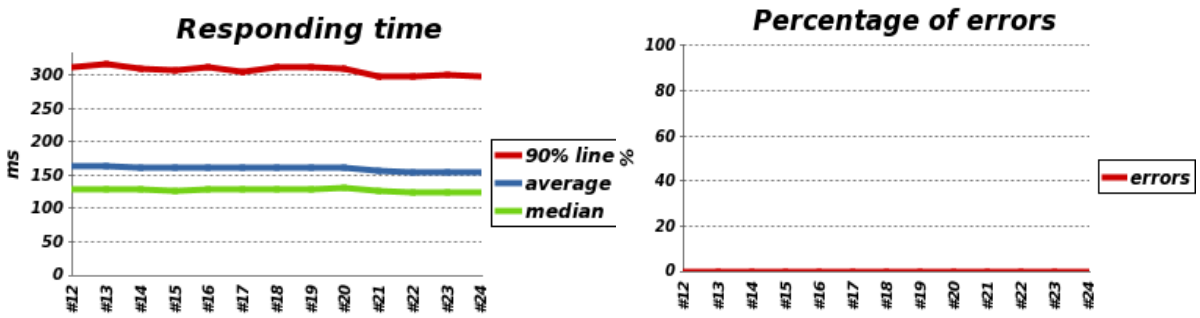


Figure 6: Performances trend

Architecture of the tests:

- Plan of tests is an xml document describing tests with a global description, an INIT part and a series of test cases, a CLEAN phase
- A test case fully describes a test. It is written in PERL and composed of 4 sections: initialisation, process, result, and finalisation.
- Gentest.pl is the tool used to generate the test plan and build the code. Another module is used to write the tests.
- A report is built every time a test plan is run. It synthesizes all results in a standard format that can be interpreted by Jenkins

8.2.2.2 For Web developments

Web developments are mainly done using PHP and Ruby on the server side; on client side, user interfaces are based on HTML (structure), CSS (presentation), and JavaScript (behavior improvement).

Libraries written in **PHP** are unitary tested: PHPUnit[2] is used to define and run unit tests. Functional testing is also defined using Selenium [3]and PHPUnit. Documentation of the code relies on phpDocumentor[4].

When implemented in **Ruby**, unit and functional tests are performed thanks to Test: Unit[5] Ruby library.

Are also used:

- simplecov[6] to have an overview of the code covered by the tests.
- Ruby-prof[7] to perform code profiling

In PHP and Ruby, JMeter[8] is used to measure applications performances.

2 http://www.phpunit.de/manual/current/en/index.html

3 http://docs.seleniumhq.org/

4 http://www.phpdoc.org

5 http://ruby-doc.org/stdlib-1.9.3/libdoc/test/unit/rdoc/Test/Unit.html

6 https://github.com/colszowka/simplecov

7 https://github.com/rdp/ruby-prof

8 http://jmeter.apache.org/

8.3 Integration level testing

This level testing is focused on the transfer of data and control across developed internal and external interfaces. For this aim it integrate above mentioned modules (image and text search engine) and it will be done as described in point 8.1 in two step approach.

8.3.1 Integration of the modules

8.3.1.1 Centralised architecture

- An image (associated data) is sent from a biobank (with the software application for sending samples) and reaches to the Centralised Index of Images.
- A sample arrives at the Centralised Index of Images and its pyramidal structure of images it's stored locally, in the Centralised Index of Images.
- A sample arrives at the Centralised Index of Images and its thumbnail image it's stored locally, in the Centralised Index of Images.
- A sample arrives at the Centralised Index of Images and the image search engine is informed. The corresponding images are provided.
- A sample sent from the Centralised Index of Images arrives to image search engine.
- A sample arrives at the Centralised Index of Images and the text search engine is informed. The associated clinical data are provided.
- Clinical data sent from the Centralised Index of Images arrive to the text search engine.

8.3.1.2 Distributed architecture

- A sample is selected in a biobank to be shared in BIOPOOL with a software application.
- This software processes the sample and the pyramidal structure of images is generated and sent from the biobank to the Centralised Index of Images.
- This software informs to a module the image search engine about a new sample arrival and the descriptors of the image are extracted and sent to the image search engine.
- This software informs to a module of the text search engine about a new sample arrival and the descriptors of the clinical data are extracted and sent to the text search engine.
- The pyramidal structure of images is stored in the Centralised Index of Images.
- The thumbnail image of a sample is stored in the Centralised Index of Images.

8.3.2 Control across the web portal

- A sample arrives at the Central Index of Images and is properly indexed in the database
- On entering the Web Portal, "the most searched samples in BIOPOOL" are automatically searched and presented in the Home.
- On entering the Web Portal, "the most viewed samples in BIOPOOL" are automatically searched and presented in the Home.
- On entering the Web Portal, the total number of pathologies is automatically searched and presented in the Home.
- On entering the Web Portal, the total number of samples is automatically searched and presented in the Home.
- From the Web Portal, samples are searched based on a sample image and matched samples are obtained.
- In the image that is entered as search criteria, a ROI (rectangular, circular) of the image can be selected to indicate the area of the image to be considered as a search.
- From the Web Portal, samples are searched based on text criteria established in the proof of concept and matched samples are obtained
- From the Web Portal, samples are searched based on text and image criteria and matched samples are obtained
- From the Web Portal, the full record of a sample found is accessed, showing a set of clinical data (like pathology) and the image(s).
- From the Web Portal, a sample can be displayed in the Web viewer.
- A sample is displayed in the Web viewer
- A sample can be navigated to any area of the image.
- A Window guide shows the currently-enabled area with the full image in the background for reference.
- A whole image of a sample can be viewed. Zoom is adjusted automatically to see the whole image.
- Frame windows with images of interest of a sample are displayed in the left bottom part of the screen. When the user selects a certain region of interest, it can be viewed at full screen.

- A sample can be viewed at different zoom levels (1.25x, 10x, 20x, 40x, 63x, 100x) and levels can be selected from a combo box located in the toolbar.
- A sample can be commented: including notes and observations about certain regions of the image.
- A sample can be measured. Users will be able to make measures such as length or area.

8.4 System level testing

This level testing demonstrates that all specified functionalities exist and that the software product is trustworthy. It verifies the as-built program's functionalities and performance with respect to the requirements for the software product as exhibited on the specified operating platform.

Thus, once the different modules are developed and ready for the integration to constitute the proof-of-concept-prototype, a second batch of tests will have to be performed with these aims:

- To validate the whole architecture of BIOPOOL system
- To validate all processes of data flows
- To guarantee that search results provided by separate modules (text and image) are preserved in this extended way.

So, in this point we refer to the description of tests described in point 8.3, since all the functionalities are validated from the WebPortal; and for the quality of the results of the Image Search Engine we refer to the same metrics explained in point 8.1.

The results obtained will be included in the deliverable D6.3. *Report on the validation results of proof-of-concept prototype*, scheduled by M12.

System level software testing addresses functional concerns and the elements of a device's software (described in point 9) that there are related to the intended use.

9 Measure indicators

9.1 Performance issues

Issue	Time
A sample is sent from a biobank of BIOPOOL network (with the interface for sending samples) and reaches to the Centralised Index of Images.	5 minutes
A sample arrives at the Centralised Index of Images and its pyramidal structure of images it's stored locally.	1 second - 5 minutes, depending on the original format.
A sample arrives at the Central Index of Images and it's properly indexed in the database. BIOPOOLID and the biobank identifier are associated internally	Immediate
A sample arrives at the Centralised Index of Images and its thumbnail image it's stored locally.	5 seconds
The image of a sample sent from the Centralised Index of Images arrives to the image search engine.	1 minute in a LAN network - 5 minutes with an ADSL connection
The clinical data of a sample sent from the Centralised Index of Images arrives to the text search engine.	Immediate in a LAN network – 10 seconds with an ADSL connection
On entering the Web Portal, "the most searched samples" are automatically searched and presented in the Home.	Less than 2 seconds
On entering the Web Portal, "the most viewed samples" are automatically searched and presented in the Home.	Less than2 seconds

On entering the Web Portal, the total number of pathologies is automatically searched and presented in the Home.	Less than 1 second
On entering the Web Portal, the total number of samples is automatically searched and presented in the Home.	Less than 1 second
From the Web Portal, samples are searched based on a sample image and matched samples are obtained.	3 minutes in LAN network (1 Gbps). 1 for image transfer (above) + 2 minutes for results retrieval Higher if network connection is not as good for transference of query sample to the search engine and retrieve results.
From the Web Portal, samples are searched based on text criteria established in the proof of concept and matched samples are obtained	Less than 1 second for samples retrieval and less than 2 seconds for display of the result page.
From the Web Portal, samples are searched based on text and image criteria and matched samples are obtained.	3 minutes in LAN network (1 Gbps). 1 for image transfer (above) + 2 minutes for results retrieval Higher if network connection is not as good for transference of query sample to the search engine and retrieve results.
From the Web Portal, the full record of a sample found is accessed, showing their associated clinical data and (s) image (s).	4 seconds
From the Web Portal, a sample can be displayed in the Web viewer.	12 seconds
A sample can be navigated to any area of	1 second

the image.	
A Window guide shows the currently-enabled area with the full image in the background for reference.	1 second
A whole image of a sample can be viewed. Zoom is adjusted automatically to see the whole image.	1 second
Frame windows with images of interest of a sample are displayed in the left bottom part of the screen. When the user selects a certain region of interest, it can be viewed at full screen.	5 second
A sample can be viewed in different zoom levels (1.25x, 10x, 20x, 40x, 63x, 100x) and levels can be selected from a combo box located in the toolbar.	1 second
A sample can be commented. Users will be able to include notes and comments about certain regions of the image.	1 second
A sample can be measured. Users will be able to make measures such as length or area.	1 second

9.2 Response stress conditions

1. 6 biobanks (or nodes of biobanks) sending samples to the Centralised Image Index.
2. 20 concurrent users searching samples based on textual criteria.
3. 20 concurrent users searching samples based on image criteria.
4. 20 concurrent users viewing samples in the Web Portal.

5. The BCII sending images and textual data to TSE and ISE while 10 users are searching samples based on textual criteria and other 10 users are searching based on image criteria, all of them at the same time.

9.3 Operation of internal/external security features

- The information (images and clinical data) travel encrypted by the VPN between the biobanks and the Centralised Image Index.
- Samples cannot be sent to the Centralised Image Index from a location that does not correspond to a biobank of BIOPOOL network.
- An invalid user cannot access to the search area in the Web Portal.
- A user cannot access the image display without the proper permissions.
- A user cannot download the displayed images.
- Only authorized requests to the BIOPOOL API will be accepted.

9.4 Effectiveness of recovery procedures

- Daily backup servers of the Centralised Image Index and the Web Portal.
- Daily backup of the Image Search Engine
- Text search engine Source files imported should be backed up as it is lighter and it enables to rebuild the index.
- The database of the Centralised Image Index is recovered with the last backup in 30 minutes.
- The Web Portal is recovered with the last backup in 30 minutes.
- Image Search engine is recovered with the last backup in 2 hours
- Text search engine is recovered with the last imported files backup by data indexation in 30 minutes

10 Pathologists (and end users) requirements' validation

The functional and technical requirements of the end users are taken into account in the elaboration of validation plan. Some of these requirements are previously described in D1.2.

10.1 Phases /Chronogram and Set of images

As described in point 8.1 of this document the BIOPOOL system will be developed, tested and validated in a two stage approach (during M12 and M24).

10.1.1 First stage validation

It will be based on the Proof-of-Concept (PoC) model of the BIOPOOL system described deeply in s D6.1). As it has been previously mentioned in point 8.1 images and associated data of colon carcinoma will be use.

After signing an outsourcing contract with Tecnalia, eMedica and Pertimm (and obtain the approval of Ethic committee in Basque Country and Netherlands) they started using these data (figures and text data: see Annex I and II) to develop both modules (text and images search engine) and define index procedure.

In order to ensure the quality of image search engine it has been accorded previous verification with pathologists in month 8 and 10. Tecnalia (together with BIOEF) will work with Basque pathologists to:

- Identify the most selective descriptors, performing similarity labeling in colon carcinoma images. There is a group of 6-7 pathologists that will collaborate in this task. Several couple of images will be shown to the pathologists and they will indicate how similar every couple of images is, indicating "very similar", "similar", "poorly similar" or "not similar at all". This knowledge will be gathered through a specifically developed platform, and will be used for the definition of the similarity metrics that the retrieval search engine will use. This will be thoroughly explained in the D4.2 – Report on image search engine, scheduled by M12.
- Establish the procedure to value how good the search result is: with this post, these outputs are considered reasonable by similarity.

Thus in Month 12 it will be done a complete validation of the search engine basing on colon carcinoma samples.

10.1.2 Second stage validation

During the second year of the project it will be included more colon carcinoma samples, lung carcinoma and breast carcinoma in the system. This pool of new images (and associated data) will be translated to a new set of descriptors and codes.

Thus it will be necessary to follow the same procedure during the end of the project period M20-M24.

10.2 Users Requirements and Parameters

The system will be able to satisfy the users' requirements already listed in D1.2. Some of them will be measured by indicators previously defined in point 9.

10.2.1 Searching criteria

Following the requirements described in D1.2 during the validation it will be necessary to fulfil all of them:

- Image search: Users can upload their own image in any format, JPEG or BMP, in the same way images are stored in the system. The system will then perform a search with the acquired morphological aspects of this uploaded image on the pools of images in BIOPOOL with associated metadata. Users get a (set of) link(s) to BIOPOOL images that are found to match with the uploaded image. These BIOPOOL images can be viewed using a viewer within the system. Associated metadata is provided along each opened image in a text format.
- Text search: Users can insert text that represents a type of tissue and/or conditions in dedicated text-entry boxes. Users can also use SNOMED entries in dropboxes. Users get a (set of) link(s) to BIOPOOL images that are found to match with the uploaded image. These BIOPOOL images can be viewed using a viewer within the system. Associated metadata is provided along each opened image in a text format.
- Image-Text search: A combine search will speed up the search procedure and to better guide the system towards proper matches on image-based morphological aspects, improving the accuracy.
- The search queries will be stored as a text file including the name of the uploaded image, text, queries, date and time.

10.2.2 Web portal Appearance

At the end of M12 it will be validate the appearance and usability of the web portal by Pathologists. Then, it will be included all the collected suggestions in order to improve it and validate in M24.

10.2.2.1 Start module:

Different functions/options may be validated:

- Show the start screen with the different options to be selected by users.
- Show Pathologist identification (if registered) or login button
- Show date and hour
- Top bar including the main menu options
- Login information
- Breadcrumb trail indication (at the top of each page)

10.2.2.2 Search module:

This module will enable advanced search of samples located in the centralised index.

There will be several types of search to be validated:

- **By text**: based on contextual semantics.
    - After introducing keying the pathology or SNOMED code it will be listed the query drop boxes.
    - After fill in the needed features, the system will provide the list of images related to this query.
- **By image**: As mentioned previously in point 10.2.1 the user will be able to upload an image to be compared to other ones included in BIOPOOL pool.
- **Combine, by text and image**: users will be also able to combine text and image search to get better and accurate results.

10.2.2.3 Visualization module

All this functions will be displayed in a tool bar located in the top of the screen.

- **Navigate**: users can move to any area of the image using the mouse controls.
- **Window guide**: to show the currently-enabled area with the full image in the background for reference.
- **View the whole image**: zoom is adjusted automatically to see the whole image.
- **Location of regions of interest**: frame windows with images of interest are displayed in the left bottom part of the screen. When the user selects a certain region of interest, it can be viewed at full screen.
- **Comments**: users will be able to include notes and comments about certain regions of the image using mouse left button. Each comment will be linked to a specific position of the image.
- **Measuring tools**: the user will be able to make measures such as length or area.
- **Different zoom levels**: 1.25x, 10x, 20x, 40x, 63x, 100x levels can be selected from a combo box located in the toolbar.
- **Export**: screen shots of the current screen view can be exported in JPEG or TIFF format.

10.2.2.4 Administration module

Only those users with administration profile will have access to this module. They will be able to manage portal features and configurations by accessing this module.

Only those users with administration profile will have access to this module. They will be able to manage portal features and configurations by accessing this module.

In this respect, the administrator profile will be able to handle the following aspects:

- Controlling the publishing of website content
- User account management: this task addresses issuing, modifying and closing user accounts and related user privileges. Logins and passwords will be assigned where

appropriate including reset passwords procedures. Rights and obligations related to access to website should be contractually arranged for all type of users. All the user account data will be properly secured.

- Monitoring web traffic
- Monitoring and tuning server performance
- Search Engine Optimization: increase visibility of the website in the results provided by search engines

10.2.2.5 Personal area "My BIOPOOL"

Registered users will be able to enter his private area after introducing user name and password. This area will show the following content:

- Personal data: it will be possible to edit
- Issued orders: order details
- Historical searches
- Annotations

10.2.3 Web portal functionalities

It will be necessary to validate all the possible actions that a pathologist will do through the web portal. These actions have been already mentioned in point 8.3.

- Load an image from file to the web portal
- Select the region of interest (ROI) the users is interested in
- The search is launched by the user
    - The organ of interest is indicated
    - The text keywords are introduced
- Search Results

- Choose the best fitting of the query, ordered from “most similar” to “less similar”
- Visualization of the selected samples in a low resolution (including a few associated data)
- The information of the Biobanks that owns the selected images.
- Refining the search

10.2.4 Technical requirements

During the development of the project different technical requirements may have been appeared. Some of them has been already described in D1.2 and will be validate in M12 and M24.

- Images with different resolutions will be possible to include and use as a search query in BIOPOOL system.
- The intensity and colour of the images (up to the minimal requirements) will not affect in the search query
- The end users satisfaction box (see D1.2 point 6) will be useful to improve the search functionality system. Users satisfaction will be based on:
    - The amount of images properly matched
    - The retrieval speed
- The amount of users logged at the same time will not disturb the system functionality.
- The web portal may have to associate correctly to the information linked to each one.
- The system will have an integrated help function that can be accessed any time
- The system will block the log in of users when the maximum amount of users allowed is exceeded (reasonable maximum to be determined)
- The system will generate an error message when improper images are uploaded: minimum – maximum size and/or resolution exceeded, wrong type of file extension, no morphological aspects found, image setting outside the accepted range
- When one or more pools of images are temporarily out of use (e.g. server on which they are hosted is down), the system will warn users when they are searching on images within these hampered pools.
- Users may be logged in for a maximum amount of time: the system will log out users when they have not used BIOPOOL for a specific duration of time (duration to be determined)

11 Bibliography

- General Principles of Software Validation; Final Guidance for Industry and FDA Staff; Document issued on: January 11, 2002 http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm085281.htm
- Glossary of Computerize system and Software development terminology http://www.fda.gov/iceci/inspections/inspectionguides/ucm074875.htm

12 References

[1] Chapelle, Olivier, et al. "Expected reciprocal rank for graded relevance." Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009.

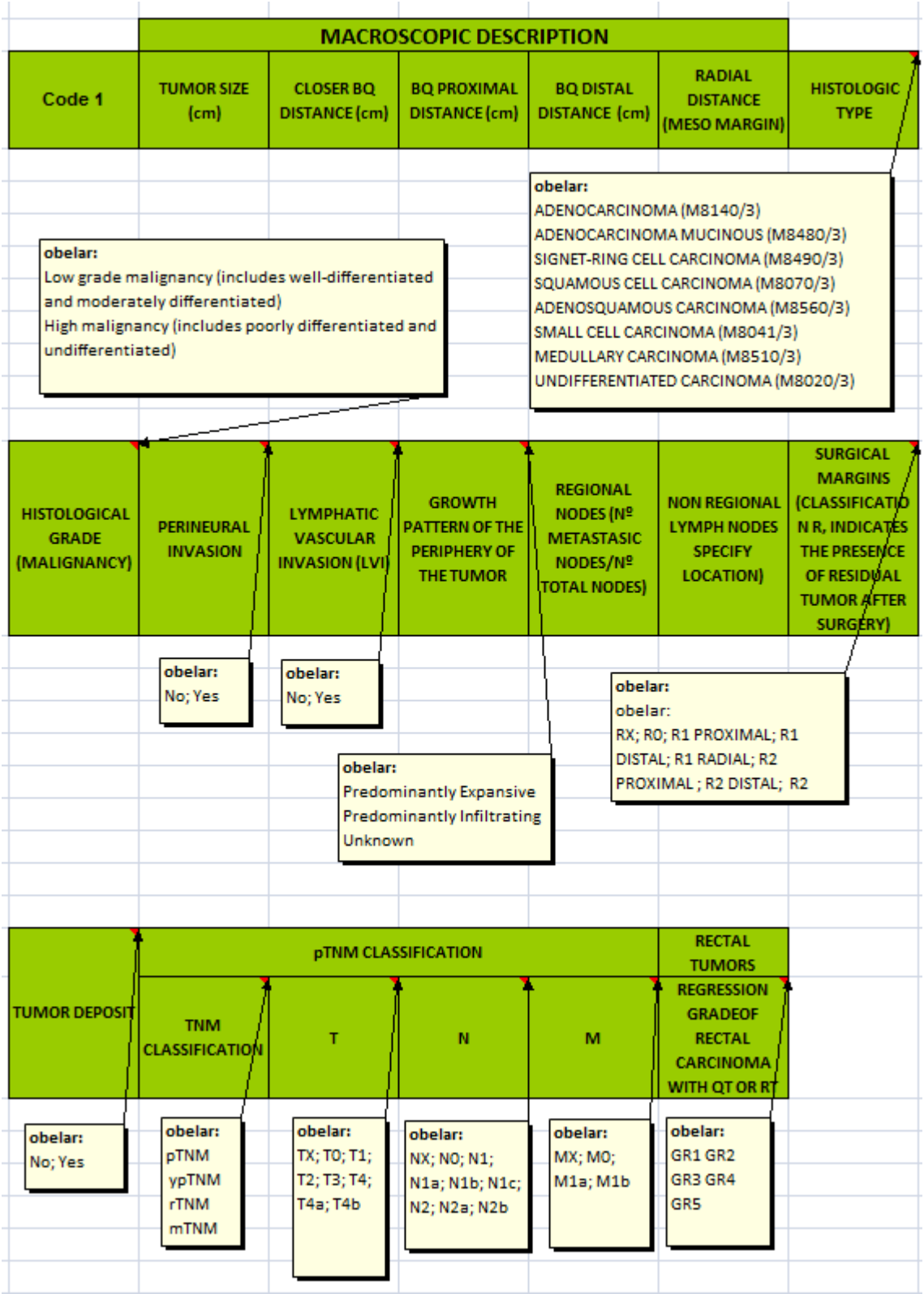
[2] Buckley, Chris, and Ellen M. Voorhees. "Retrieval evaluation with incomplete information." Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2004.

[3] Büttcher, Stefan, et al. "Reliable information retrieval evaluation with incomplete and biased judgements." Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007.

[4] Jan De Beer and Marie-Francine Moens. 2006. Rpref: a generalization of Bpref towards graded relevance judgments. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '06). ACM, New York, NY, USA, 637-638.

13 Annexes

13.1 Annex I: Associated Anatomic Pathological data (colon carcinoma)



13.2 Annex II: Associated histological pattern data (colon carcinoma)

Adyacent healthy tissue:

code 1	STRAIGHT AND UNIFORM GLANDS	EQUIDISTANT GLANDS	GOBLET CELLS WITH MUCUS	CRYPTS	UNIFORM INTERGLANDULAR SPACES	REGULAR, PERIPHERAL, BASAL GANGLIA

REGULAR SIZE	EVENLY DISTRIBUTED STROMAL	PANNET CELLS	ENLARGED NUCLEI; HYPERCHOMATIC IRREGULAR CONTOURS	MACRONUCLEOLI	FREQUENT MITOSES	ATYPICAL MITOSIS

Tumor sample:

code 1	STRAIGHT AND UNIFORM GLANDS	EQUIDISTANT GLANDS	GOBLET CELLS WITH MUCUS	CRYPTS	UNIFORM INTERGLANDULAR SPACES	REGULAR, PERIPHERAL, BASAL GANGLIA

REGULAR SIZE	EVENLY DISTRIBUTED STROMAL	PANNET CELLS	ENLARGED NUCLEI; HYPERCHOMATIC IRREGULAR CONTOURS	MACRONUCLEOLI	FREQUENT MITOSES	ATYPICAL MITOSIS

All are Yes or No answers.