Grant Agreement Number **ECP 2006 DILI 510049**

ENRICH

Evaluation Report

Deliverable number	***D-7.2***
Dissemination level	*Public*
Delivery date	*15th December*
Status	*Final*
Author(s)	*Dr. Nerutė Kligienė, Institute of Mathematics and Informatics, LITHUANIA*

e*Content*plus

This project is funded under the *e*Content*plus* programme,
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

Document Version Control

Version	Date	Change Made (and if appropriate reason for change)	Name and Institution of Commentator(s) or Author(s)
01 (draft)	November 2009	The main quality evaluation results in WP3, WP4, WP5, WP5n, WP6, criteria and categories described, the illustrating diagrams in Fig.1 – 26 provided, the questions for evaluation and evaluation data given in Annexes 1. (a) – (e).	Institute of Mathematics and Informatics, Lithuania Nerute Kligiene, IMI
1.0 Final	8. 12. 2009	Conclusions and Recommendations extended	Nerute Kligiene, IMI

Document Review

Reviewer	Institution	Date and result of Review
Jakub Heller	CCP	2. 12. 2009, Recommended extension of Conclusions and Recommendations
Tomáš Psohlavec	AIP	14. 12. 2009, Approved for submission

Document Signature/Approval:

Before the table of content each document is to contain an approval signoff form.

Approved By (signature)	Date

Accepted by at European Commission (signature)	Date

1 Structure Summary

The Work Package 7 of the ENRICH project is devoted to test and validate the tools, platforms and applications developed in the project in order to evaluate their accessibility, usability and adaptability. Existing research results and widely known publications on quality evaluation from the last few years have been studied and summarized in Methodology (D-7.1) in order to select those basic principles and evaluation criteria which fulfil the ENRICH project needs. A set of principles and quality criteria has been chosen and adapted for evaluation and testing of e-applications developed in the frame of ENRICH project in Methodology created as D.7.1. The missing investigation on the reliable statistical inference – confidence intervals for the proposed quality estimators were created, used for evaluation and published as the scientific results of the project.

The sections 3.1 – 3.3 of this document describe in short the updated quality criteria and sub criteria to be used in evaluating the separate WPs, as well as the statistics and metrics used for evaluation. The sections 3.3.1–3.3.4 includes the detailed description of the evaluation criteria adapted to the work being done in the project.

The most important part of this Report is **the section 3.4** where the results of quality evaluation performed in the project are analyzed in many aspects and presented in various cuts, illustrated by numerous diagrams and figures with the corresponding comments.

The important additional materials are included into several Annexes composed as follows.

Annex 1: The Questionnaires and results obtained in testing and evaluating:

Annex 1-(a). The Questionnaire for System Usability Score applied to ENRICH project web site usability evaluation as a prototype.

The Questionnaires-interactive forms, used at the evaluation web site [9] for evaluating results achieved while working on the packages and the statistical data obtained:

Annex 1-(b) – WP 3*;*
Annex 1-(c) – WP 4;
Annex 1-(d) – WP 5;
Annex 1-(e) – WP 6.

Partner responsible for WP 7 in ENRICH Consortium:
CR 8– **Institute of Mathematics and Informatics, Lithuania**

Contract Start Date: 01.12.2007. Project duration: 24 months

CONTENT

2 Introduction

The objective of ENRICH project is to create a base for the European digital library of cultural heritage, a real research environment built upon the existing *Manuscriptorium* platform adapted to the needs of organisations holding repositories of manuscripts. The *Manuscriptorium* [1] Digital Library represents manuscripts, rare old printed books, and other documentary heritage and is used as a working field of the project ENRICH [2].
ENRICH enables research activities in a particular type of cultural repositories – manuscripts – and follows a vertical approach to enable more types of cultural organizations to integrate their repositories also into TEL – The European Library. ENRICH enables real seamless metadata and image data incorporation from dispersed resources onto a single platform with a uniform interface that will co-operate in real time with remote source data including various image banks stored in original locations of content owners. As a first step in evaluation the methodology for evaluation, testing and validation of the e-applications was delivered (D-7.1) at the fifth month (M5) of the ENRICH project work and then implemented from M9 to M23. This report describes and summarizes the results obtained in testing and evaluating activities.

The main objectives of this document are:

- To apply the created methodology (D-7.1) for evaluation of usability and adaptability of tools, platforms and applications developed in ENRICH project, with necessary changes and more exact evaluation target points according the results achieved in the project.

- To describe detailed results obtained while estimating WPs, the specified Quality Criteria and measures applied for the evaluation of e-applications developed in the framework of ENRICH.

There are many diverse and interdependent tasks to be resolved and implemented in the project. The features and attributes of the results of these tasks – tools, applications and processes – were more precisely defined as the ENRICH proceeded. The particular features of each object of evaluation were defined taking into account the end-users reviews performed during the project, their subsequent analyses, their implementation, and also the implementation of preceding interdependent task results.

The evaluation itself was performed based on the detailed documentation as soon as each task and subtask produces particular results – so an instant feedback could be provided to the Task and WP leaders enabling them, in case it is needed, to take correction steps.

The process of creating of the detailed documentation for evaluation was run as follows:

- The necessary information on subject of evaluation and questions of the survey designed in cooperation of particular task leader for each task and the WP7 leader, where an evaluation was planned.
- The task leader proposes particular content and form of the survey, the WP7 leader ensures it is in conformance with methodology.
- Subsequently the survey has been discussed with participating partners and the agreed corrections implemented.

- Final form of the survey published on-line at the special evaluation website [9]; gathering of required responses is ensured by common work of participating partners.
- The monitoring of each evaluation undertaken, processing of the gathered information were undertaken by the WP 7 leader.
- The results of every evaluation performed have been summarized and published at the evaluation web site [9].

In case if there are specialised tools available to perform particular task evaluation then these tools (for example, M-Tool, SYSTRAN Translation interface) were used in order to increase the efficiency of evaluation.
The main duties in evaluation, testing and validation are described in Work Package 7 (WP 7, Description of Work).

WP-7 Work package Description					
Work package number :	**WP - 7**	**Start date:**	**3th month**	**End date:**	**23th month**
Work package title:	**EVALUATION, TESTING AND VALIDATION**				
Objectives					
▪ Evaluation of usability and adaptability of tools, platforms and applications developed while implementing the ENRICH project, such as personalization for contributors and users, multilingual and user friendly access; ▪ To test the possibilities of application of modern tools for automated translation tools for multilingual search engine over existing data and metadata and the new data sets in order to fix possible shortcomings and improve the results before ending the project.					
Description of work					
Work package leader: IMI **Task 7.1 – Defining evaluation strategy** Task leader: IMI **Task participants: NKP, AIP, OUCS, KU, BNCF, MICF, VUL, SYS, ULW, SAM, UZK, DSP, NULI, BNE, BUTE, PSNC** The basic principles and evaluation criteria developed by worldwide known teams (such as W3C, Minerva Technical Guidelines) will be studied and adopted for evaluation and testing of e-applications planed to be developed in the frame of the project ENRICH. The respective weighting of criteria for evaluation of the results should be proposed, indicating more and less important categories. Describing positives and negatives of the applications usable for improving the working methods. Evaluating the platform, allowing testing of newly processed data (i.e. automated translation) and their usefulness to new implemented or modified tools. **Task 7.2 – Testing and evaluating the accessibility, usability and adaptability of developed applications** Task leader: IMI **Task participants: NKP, AIP, OUCS, KU, BNCF, MICF, VUL, SYS, ULW, SAM, UZK, DSP, NULI, BNE, BUTE, PSNC**					

Evaluation step by step the technical aspects and usability of the system. Consortium will prepare and assess evaluation tests, following recognized usability procedures. The usability evaluation will cover the assessment of all aspects of the service and language groupings, and be carried out in partnership with the user partners. The evaluation results will be fed back to the technical partners (who will adjust the technology platform). The evaluation process will reflect all strong and weak points of the results derived in the frame of ENRICH project: interactivity, interoperability – developed system being able to share information across databases and other online entities. Tested if similar data models and metadata element sets are used for semantically similar items and concepts.

Results:

- All the e-applications developed in the project will be tested and evaluated on different content
- Problems and bugs of developed tools will be fixed and adjusted by technical partners
- All the strong and weak points of the results of ENRICH will be described in **the evaluation report**

(Inter-) Dependencies, milestones and expected results

Milestone 7.1. Tested and evaluated usability and adaptability of the e-applications (personalization contributors and users, multilingual and user friendly access, 23rd month of the project work).

The WP depends on the results of all technological WPs – WP3, WP4, WP5 and WP6.

Expected results – well tested system, contributor & user-friendly, accessible and usable in wider European community, reflecting wider cultural requirements, ensuring the content to be perceivable, operable and understandable by the broadest possible range of users and compatible with their wide range of assistive technologies, now and in the future.

Deliverables

D7.1. Methodology for Testing and Validation of e-Applications (m5); responsible partner: IMI.

D7.2. **The Evaluation Report** (m23); responsible partner: IMI.

The five items were decided to be evaluated using the Methodology for Testing and Validation of e-Applications. Those items are:

- **The usability of the collaborative environment in the ENRICH project** (the pilot evaluation performed up to May 2008);
- **Migration Tool developed in WP 3** (started in April 2009 and finished in the middle of May);
- **Personalized Translation Interface developed in WP 6** (started in April 2009 and was finished in the middle of May)

- **Possibilities for sharing of large data sets investigated and developed in WP 5** (evaluated during 15 May – August 2009, New facilities evaluated October – November 2009);
- **Tools for creation of virtual documents by researchers developed in WP 4** (evaluated September 2009);

The evaluation added more value to the whole ENRICH project outcomes, it provided a comparison and the necessary feedback to the project Content (and Associated) Partners and last but not least: it complies with the important trends observed by the prominent European projects and their results. Additionally, a quality of submissions to *Manuscriptorium* was improved considerably as described below.

2.1. Quality Evaluation of Digital Documents Provided in Manuscriptorium

In addition to above listed project results to be tested and evaluated, evaluation of quality of digital documents provided in *Manuscriptorium* was proposed. During the ENRICH project, various digital documents are provided in *Manuscriptorium*. The documents come from many places, and are created under different local conditions, using various procedures. A very important criterion for user is usability of image – the visual perception. The metadata, which carry the information about the original documents and also provide the connection to the digital images, are on same level of importance. In the long time perspective of the use of such a data, specification of other features of digital data, and also evaluation of the conditions of their creation is important.
When digitizing in order to create an access to cultural heritage, it is necessary to consider the following aspects:

- unique value of the document,
- safety of the digitization process for the original,
- quality of digital data,
- quality of metadata,
- safety and reliability of data preservation,
- usability, accessibility (formats, metadata).

When specifying these aspects, the goal is to ensure well-balanced digital research environment. The motive of the quality evaluation is not to promote any kind of comparison, or rivalry among the partners. For example, it is possible, that the only and therefore very important image of some document just exists, taken by camera built in mobile phone, and therefore, this image is acceptable in *Manuscriptorium*.
The detailed methodology “Technical Quality Criteria” were prepared by the technological coordinator AIP and issued as a separate document by AIP.

3 The Quality Criteria for Validating ENRICH Work

The evaluation of the ENRICH results is the core activity of the WP 7. Existing research results and widely known quality publications from the last few years have been studied by the partners to select those basic principles and evaluation criteria which fulfil ENRICH project needs.

3.1. User-Based Objectives. The Set of Quality Criteria

According to the MINERVA Technical Guidelines [3] the important areas for consideration at least have to include:

Interoperability: It is important that content can be accessed seamlessly by users, across projects and across different funding programmes. It should be possible to discover and interact with content in consistent ways, to use content easily without special tools, and to manage it effectively.
Accessibility: It is important that materials are as accessible as possible and are made publicly available using open standards and non-proprietary formats. If material is to be a widely useful resource it will be necessary to consider support for multiple language communities and ensure accessibility for citizens with a range of disabilities.
Preservation: It is important to secure the long-term future of materials, so that the benefit of the investment is maximized, and the cultural record is maintained in its historical continuity and media diversity.
Security: In a network age it is important that the identity of content and projects (and, where required, of users) is established; that intellectual property rights and privacy are protected; and that the integrity and authenticity of resources can be determined.

ENRICH Consortium prepared and implemented evaluation tests, following defined usability procedures. The usability evaluation covered all aspects of the service and language groupings, and were carried out in partnership with the user partners. The evaluation process reflected all strong and weak points of the results developed in ENRICH project: interactivity, interoperability – so that the system is able to share information across databases and other online entities. The evaluation process include the selection of the main criteria and sub criteria and then their deployment. The evaluation procedure contains also description of positive and negative features of the applications, which have to be used to improve the working methods. The evaluation results do serve as the feedback to the technical partners (helping to adjust the technology platform accordingly).

The main idea developed for ENRICH evaluation in the Methodology was to select a reasonable number of background principles expressed in the Criteria Set, each Criterion having a number of sub criteria (their maximal number was fixed up to 12). The sub criteria were classified into categories, reflecting the different tasks to be tested. Each category has a fixed position in a set of sub criteria. For example, the first three criteria refer a digital object, the next three the software tools used, another three the processing properties and so on. If a smaller number of sub criteria are sufficient in some category, the numeration in the next category should be started from the number fixed to that category. Such structured model allows to test and to evaluate against the same universal principles multifaceted items:

separate digital objects, processes or tools**.** Similar approach was used for modification of a Usability test to be done by users having a general interest.

A limited number – five criteria covering the most important aspects of digital repository functionality has been selected and is listed below shortly. Each criterion can have maximum 12 sub criteria, all together we can consider up to 60 features, reflecting various attributes of multifaceted objects under investigation for testing of the tools/process/objects developed in WP3, WP4, WP5, WP6. In preliminary version the categories of tools in Adaptability and Usability had a number of sub criteria equal to 9 and 11, correspondingly. Later on when a real content for testing was available in ENRICH project, the list of sub-criteria was updated and improved. But the same structured model and categories were kept. We have considered the following categories:

i. the separate items – digital **objects** for submission or delivery,
ii. the software **tools** used,
iii. **processing** of assessment, ease of deployment,
iv. User-friendliness and long term preservation properties in a level of whole **repository**.

The set of sub criteria is classified into categories according to (i) – (iv) statements in each of the 5 main Criteria. The following structure is proposed to use in each *Nth* CRITERION (*N* – numerated from 1 to 5).

N. **CRITERION (an explanation of proposed structure in each of CRITERIA)**
N.1. sub criterion concerning a **digital object** – separate item
N.2. sub criterion concerning a **digital object** – separate item
N.3. sub criterion concerning a **digital object** – separate item
N.4. sub criterion concerning a separate **tool/ software** tested
N.5. sub criterion concerning a separate **tool/ software** tested
N.6. sub criterion concerning a separate **tool/ software** tested
N.7. sub criterion concerning **process/activities** tested
N.8. sub criterion concerning **process/activities** tested
N.9. sub criterion concerning **process/activities** tested
N.10. sub criterion concerning **whole repository**
N.11. sub criterion concerning **whole repository**
N.12. sub criterion concerning **whole repository**

It was decided to use the following Main Criteria each containing up to 12 sub criteria structured as explained above. Therefore the double numeration of our sub criteria was used as *N, j* (*N*= 1, 2,…, 5; *j* = 1, 2,…, 12). The list of criteria is as follows:

1. **INTEROPERABILITY** (with up to 12 sub criteria specified by experts)
2. **ADAPTABILITY** (with up to 12 sub criteria specified by experts)
3. **USABILITY** (with up to 12 sub criteria specified by experts)
4. **SECURITY AND PRESERVATION** (with up to 12 sub criteria specified by experts)
5. **MULTILINGUALITY** (with up to 12 sub criteria selected by experts)

This is a general set of criteria containing up to 60 structured sub criteria. In each WP evaluation case the sub criteria were specifically determined together with each WP leader and are described later when presenting the WP3, WP4, WP5, WP6 evaluation results in the 3.4 section of this Report.

3.2. Metrics for Evaluation of Quality, Prototype of Evaluation and Problems of Reliable Statistical Inference Based on Evaluation Results

Each sub criteria was decided to rate using a range of **0 – 4.** These ratings defined as:

0 – Failed or feature does not exist,
1 – Has poor support and/or it can be done but with significant effort,
2 – Fair support but needs modification to reach the desired level of support,
3 – Good support and needs a minimal amount of effort,
4 – Excellent support and meets the criteria out of the box, minimal effort.

The questionnaire proposed in [4] had also five degrees from „*Strongly disagree*" to „*Strongly agree*" which later arranged to have a range from 0 to 4. It has been extensively used and adapted in usability testing since 1986 and is the most strongly recommended of all the public domain questionnaires, according the *UsabilityNet.org* experts [5]. The classification of sub criteria according the categories used for the ENRICH purposes were made additionally in it and used as a model for estimating Usability of the ENRICH collaborative environment. The questionnaire modified from [4] is added separately in the Annex 1-a. The importance rating for sub criteria, proposed in some evaluation methodologies, was not applied in our Methodology because of a real risk to distort evaluation results by using non-objective importance rating.

The different kinds of users from target audience were identified as:

- Content, information provider, manager
- Technical staff, supporting personnel
- Scholar, researcher in the historical documents area
- End-user with a general interest.

In reality the full list of sub criteria can be hardly assessed for evaluation by each type of target user, therefore it is natural to tolerate omitting a few of them. For example, content or information provider is strong on metadata submission and repository maters, while technical personnel can focus on processing and interoperability of platforms. The statistics for evaluation is derived as a double average value over users and sub criteria used – an average score of some criterion expressed as $\hat{e} = 1/kn\sum(assigned\ scores)$, where n is the sample size – the number of respondents, k – the number of sub criteria used. The average estimate $\hat{e}$ give us more reliable result than could do the individual estimates of experts summed, often used to evaluate repository. It is recommended to fix responses in focus groups (estimates by contributors, technical staff, and researches) separately. Evidently, as more users are involved in each focus group, the more reliable results can be derived. The expected size of quantity nk is around 50 – 70 (number of users involved into evaluation – respondents, multiplied by the number of questions answered). This quantity is satisfactory to get a reliable evaluation of any criterion under study.

The statistical background for developing convenient statistics for quality evaluation enabling to derive reliable results independent of a number of criteria / sub criteria used for evaluation was elaborated and presented for scientific community at two well known international conferences [6, 7]. The articles were accepted and published as valuable scientific results. Those two investigations – published scientific articles are added in the Annex 2.

3.2.1. System Usability Score Applied to ENRICH Project Web Site. A Prototype of Evaluation

Evaluation of quality of e-Applications developed in the Framework of ENRICH is not simple It can be performed mostly by experts and validated only when the corresponding e-Application/service/tool are fully accessible to users. An alternative simple method was used to evaluate usability of collaborative environment in ENRICH project web site [2]. The questionnaire has been published online [9] and partners were asked first to fill it in when accessing the project web site. The users, classified into four target groups above, were asked to tick the appropriate box indicating that they have previous experiences in the area as:

- Content provider, information manager ☐
- Technical personnel, supporting staff ☐
- Scholar, researcher in historical documents ☐
- End-user having general interests ☐

This classification enables developers to investigate the usability aspects in context of separate target groups.

Using System Usability Scale (SUS)

The system usability scale is generally used after the respondent has had an opportunity to use the system being evaluated, but before any debriefing or discussion takes place. Respondents should be asked to record their immediate response to each item, rather than thinking about items for a long time.
SUS yields a single number representing a composite measure of the overall usability of the system being studied. Note that scores for individual items are not meaningful on their own. Statistical inference on the system usability will be made only after the statistical data collected and processed.
Using SUS all items should be checked. If a respondent feels that he/she cannot respond to a particular item, they should mark the centre point of the scale. In the alternative, the more complicated evaluation, developed in the section 3.3, the respondent has possibility to answer questions selectively, according the areas of his best knowledge.

The Results of Pilot Evaluation

From the first 34 respondents evaluating the quality of ENRICH working environment – answering 10 questions provided in the form online www.musicalia.lt/sus/ [9] (the results received up to 13 May 2008 were counted) and added to this report in Annex 1-a. We can conclude that in average different kind of users (content providers, information managers, technical personnel, supporting staff, scholars, researchers in historical documents, end-users having general interests) expressed rather similar opinion about the usability of project web site (Fig.1). Surprisingly, all kinds of users had an average score of usability of this site rather evenly distributed (Fig.2). Processing was ranked highest by all users and the tools received the lower evaluation (Fig.3). But the total number of respondents was rather small to confirm such conclusion with statistical confidence.

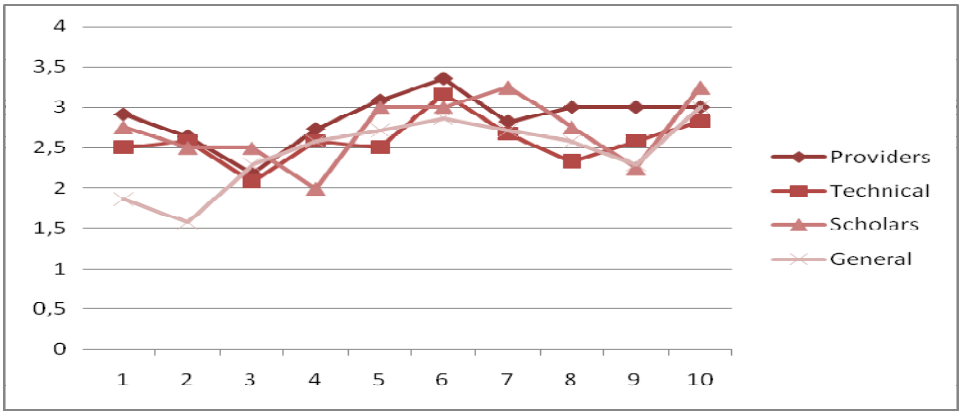


Fig.1. The average evaluation of 10 questions by different kinds of users.

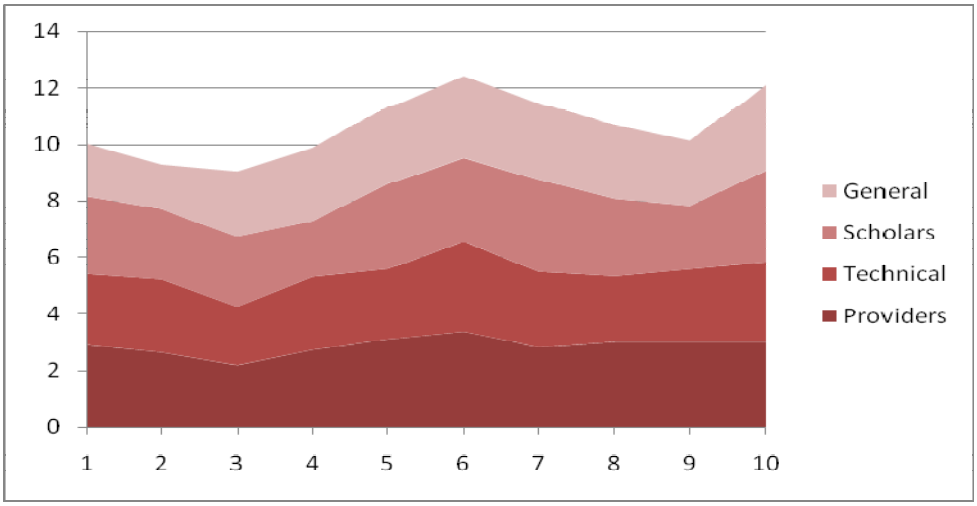


Fig.2. All kinds of users had rather similar opinions on usability of the ENRICH web site.

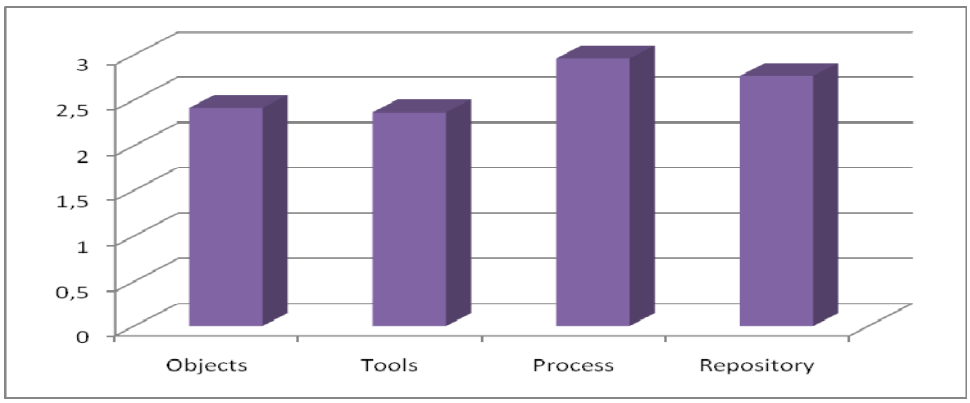


Fig.3. The average of points assigned by respondents to the categories.

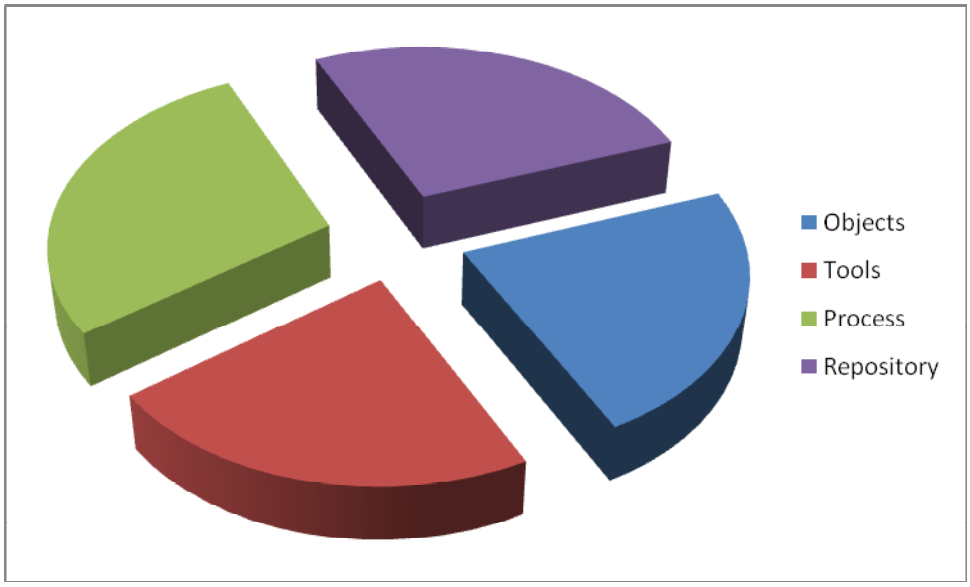


Fig.4. The usability composed of four categories and evaluated by 34 respondents reached the summary value 10,51 (from the maximum possible 16 score) or 65,69 %.

The components of averaged categories have rather equal impacts to the total as seen from the numerical values provided below:

Objects	2,41375
Tools	2,36625
Process	2,96125
Repository	2,76875
Usability	10,5100

Using the agreed ratio to the maximal possible value, the standard **Usability** score can be evaluated by its percentage (10,51/16) 100 = 65,69 i.e. it is falling into interval 26 – 75, that is rather good. By standardized scores of categories we would have the following scores in percentages:

Objects	60,28
Tools	59,16
Process	74,03
Repository	69,22

The conclusion from the first and not large sample of respondents was positive: the usability of web site as collaboration environment of partners **is rather good** (the score is 65). The same is applicable to all categories: objects, tools, process, and repository. Although the difference in scores of categories less than 3 percents are not significant due to moderate sample size (equal to 34) – this inference come from more sophisticated statistical investigation of sample means properties [6, 7]. Therefore the difference between objects and tools is negligible, but it is significant in a case of process and repository. This pilot evaluation was made in order to demonstrate that proposed method works on a system

investigated and we are able to derive numerical evaluations of quality and interpret them correctly during the extended evaluation of project results.

3.2.2. Creating Statistics for Reliable Statistical Inference Based on Evaluation Results

Here the statistical background of a reliable quality evaluation is described in short. The binomial distribution is a natural model for series of trials with possible two outcomes, in our case they are the responses "*Yes*" and "*No*" of users expressing their satisfaction in digital environment, implemented in the evaluation site [9]. Each property under evaluation has 4 questions to answer "*Yes*" or "*No*" resulting in a score from 0 to 4. Let p denote the probability of "*Yes*" on a single trial and let n independent trials with a constant probability p were accomplished. Let n respondents of one of a target audience cluster, specified above, have to evaluate the quality using k sub criteria when the result of individual evaluation should be ranging from 0 to 4. Denote X_{ij} – the random variable representing the opinion of i-th respondent ($i = 1, 2,\ldots, n$) on j-th sub criterion ($j = 1, 2,\ldots, k$). The possible values of X_{ij} are the integer values: 0, 1, 2, 3, 4. Let us define additionally the indicator random variables $Y_{ij}^{(r)}$ ($r = 1, 2, 3, 4$) as the independent components of each X_{ij} , such that

$$X_{ij} = Y_{ij}^{(1)} + Y_{ij}^{(2)} + Y_{ij}^{(3)} + Y_{ij}^{(4)}, \text{ where} \quad (1)$$

$$Y_{ij}^{(r)} = \{1, \text{ if the answer is "}Yes\text{" (with probability } p), \\ 0 - \text{otherwise (with probability } 1-p)\} \quad (2)$$

Then the mean $E\{X_{ij}\}$ and the variance $V\{X_{ij}\}$ of those random variables can be calculated using the well known properties of the binomial distribution:

$$E\{X_{ij}\} = 4p;\ V\{X_{ij}\} = 4p(1-p). \quad (3)$$

Let us construct the statistics for quality evaluation as a double average over responses of n respondents using k sub criteria as

$$\boldsymbol{X} = 1/nk \sum_{i=1,\, j=1}^{n,\, k} X_{ij,} \quad (4)$$

The Eq. (4) represents an estimator $\boldsymbol{X}$ of some quality principle (Criterion, item, property) when the average is derived using k sub criteria and estimated by n respondents. The mean and the variance of a random variable $\boldsymbol{X}$ are respectively $E\{\boldsymbol{X}\} = 4p$ and $V\{\boldsymbol{X}\} = 4p(1-p)/nk$. The maximum possible value of $\boldsymbol{X}$ is $max\ \boldsymbol{X} = 4nk/nk = 4$. Then statistic suitable for our investigation is the ratio: $\boldsymbol{X} / max\ \boldsymbol{X} = \boldsymbol{X} / 4$, representing a proportion to the maximal value. The distribution of $\boldsymbol{X}$ is well approximated by the normal law for $4nkp > 5$ and $4nk(1-p) > 5$. Based on this approximation we will construct the confidence intervals for investigated proportions of users having one or another opinion on the specified question under investigation.

Let us consider the standardized version of $\boldsymbol{X} / 4$ (i.e. minus its mean value then divide by its standard deviation) and use the large-sample approximation to that statistic by normal law. If we denote by $\boldsymbol{p}^* = \boldsymbol{X} / 4$ the proportion of observations in a sample of size n that reflects an opinion of n respondents on a quality of interest, estimated by using k sub criteria, then an approximate $(1-\alpha)$ confidence interval for the unknown proportion p of the population that have such opinion is

$$\boldsymbol{p}^{*} - z_{\alpha/2}\,[p\,(1\text{-}p)/\,4nk]^{1/2} \le p \le \boldsymbol{p}^{*} + z_{\alpha/2}\,[p\,(1\text{-}p)/\,4nk\,]^{1/2}\,, \qquad (5)$$

where $z_{\alpha/2}$ is the upper α/2 percentage point of the standard normal distribution. Replace unknown p in the upper and lower limits in Eq. (5) by its estimate $\boldsymbol{p}^{*}$ and denote $\Delta = z_{\alpha/2}\,[\boldsymbol{p}^{*}\,(1\text{-}\,\boldsymbol{p}^{*})\,/\,4nk]^{1/2}$. For 0,95 confidence interval $z_{\alpha/2} = 1{,}96$ and from the Eq. (5) it follows that the upper and lower confidence limits are:

$$\boldsymbol{p}^{*} \pm 1{,}96\,[\boldsymbol{p}^{*}\,(1\text{-}\,\boldsymbol{p}^{*})/4nk]^{1/2} \quad = \boldsymbol{p}^{*} \pm \Delta \qquad (6)$$

Here $\boldsymbol{p}^{*}$ is the estimated proportion, ranging from 0 to 1. Sometimes it is more natural to consider percentages instead of proportions. Then let $\boldsymbol{p}^{*}\% = 100\,\boldsymbol{p}^{*}$ and $\Delta\,\% = 100\,\Delta$, respectively. Let $K = 4nk$ and $\Delta\,\%$ be corresponding to 0,95 confidence level. The relationships of Δ, K, and $\boldsymbol{p}^{*}$ are presented in the *Table 1*. Analyzing the variability of Δ with respect of $\boldsymbol{p}^{*}$ we conclude that the confidence interval will always have a maximum for $\boldsymbol{p}^{*} =$ 0,5 and will have minimum values symmetrically at the ends of its range – for small and large $\boldsymbol{p}^{*}$ values, as clearly seen from the numerical values in the *Table 1*.

Applying the formulae (6) for rather large K, say $K = 500$, we will get $\boldsymbol{p}^{*} \pm 0{,}026$ (in case of $\boldsymbol{p}^{*} = 0{,}1$ or 0,9) and $\boldsymbol{p}^{*} \pm 0{,}044$ (at the $\boldsymbol{p}^{*} = 0{,}5$). This allows us to state with 0,95 confidence level that if $4nk$ is close to 500, the real percentages can vary from derived estimate approximately from ± 2,6% to ± 4,4%. Notice that if we have rather moderate sample size, say $n = 30$, but we can use at least 4 sub criteria ($k = 4$) then the approximate 0,95 confidence limits are similar to those calculated for $K = 500$ and provide us with the satisfactory result having variation around 3%. So the variability can be decreased not only by increasing sample size but also by involving more sub criteria for evaluation. This conclusion is important for implementation when larger sampling is too expensive or sometimes even impossible.

Table 1. Δ% dependence on $K = 4nk$ and $\boldsymbol{p}^{*}$

Observed $\boldsymbol{p}^{*}\%$	10% or 90%	20% or 80%	30% or 70%	40% or 60%	50%
Δ % if $K = 100$	5,8800	7,8400	8,9818	9,6020	9,8000
Δ % if $K = 500$	2,6296	3,5061	4,0168	4,2941	4,3827
Δ % if $K = 1000$	1,8594	2,4792	2,8403	3,0364	3,0990

The Eq. (6) can be used to calculate what value of sample size n is needed to attain the desired accuracy: fixing the value of Δ, the value n calculated from Eq. (6) is

$$n \ge (z_{\alpha/2})^2\,\boldsymbol{p}^{*}(1\text{-}\,\boldsymbol{p}^{*})\,/(4k\,\Delta^2). \qquad (7)$$

The variability of estimated percentages has to be taken into account when making statistical inference on digital repository quality's characteristics we are seeking to investigate. The estimator $\boldsymbol{X}$ defined by Eq. (4) gives more reliable results than individual estimates of experts. The clustered samples of users (say n_1 contributors, n_2 technical staff, n_3 researchers, and n_4 general users) are recommended to keep over whole estimation process and to investigate how the needs of each kind of the user specified are satisfied if the total sample size is large enough. Based on those sampling data it is possible to derive results

showing the weak and strong spots across the criteria or categories in context of every group of users. This enables the developers of system to fix and to improve a quality of digital repository.

3.3. Evaluation, Testing and Validation Specifically for ENRICH WPs

The four technological work packages: WP3, WP4, WP5, and WP6 are closely related to each other and interdependent with evaluation, testing and validation. Keeping the same numeration of the Tasks as described in project *DoW* under work packages, here we reformulate the items (assigning each by a triple number: for example the item 4.3.2 is #2 in the Task 4.3) which can be tested with the quality criteria described in section 3.1 and evaluated applying the metrics introduced in the section 3.2.

3.3.1. Standardization of Shared Metadata – WP 3

Objectives

To ensure interoperability of the metadata used to describe all the shared resources by analyzing the various standards used by different partners and ensuring their mapping to a single common format, which will be expressed in a way conformant with current standards.

Work package leader: OUCS

The point of the WP3 task 3.1 is to investigate and facilitate the migration from MASTER (TEI P4-based) manuscript descriptions to the ENRICH Specification (TEI P5-based). Much of this migration has already been completed and the outputs from this project are some basic XSLT stylesheets useful for migration, and a report on the development of migration tools. One of the recommendations is that these need to be customised to cope with the local encoding at any particular archive. Evaluation of this task, therefore, is simply to confirm that the stylesheets have been produced, that a report on recommended methods of migration has been created as a deliverable, and that these are freely available on the web.

The aims of WP 7 – to test and evaluate how these objectives attained.

Task 3.1: Conversion between TEI P4 and TEI P5 platforms for description of manuscripts (m2-m15)

3.1. 1. Development of migration tools on the basis of the sample data sets and validation with respect to:

a) Interoperability/Tools,

b) Adaptability/Tools,

c) Usability/Tools.

The evaluation website [9] contains the pages http://www.musicalia.lt/eta/wp3.php having all necessary or additional for evaluation information for respondents. Each of those three questions: 3.1.1 (a), 3.1.1 (b), 3.1.1 (c) has four detailed statements requiring from a respondent only to choose "*Yes*" or "*No*" and finally resulting in the score between 0 and 4 for each question corresponding to sub criteria. The questions are displayed at the evaluation web site [9] and added to this report in the Annex 1-b.

3.3.2. User Personalization – WP 4

Objectives

Enable to subdivide the contents of Manuscriptorium into thematic collections. To satisfy the needs of all Manuscriptorium end-users, thematic collections will be created and maintained by authorized experts.

Furthermore, end-users will be able to construct their own individual collections and virtual documents via usage of newly developed tools – this will create opportunities to build individual user virtual libraries according to their personal needs (such as study, teaching etc.). The tools will enable to decompose the digitized documents into necessary chunks/analytical digital objects and recompose them in new virtual documents following special teaching or learning goals, e.g. showing all illuminations from one scriptorium in a virtual document in spite of the fact that they are from various originals owned by different institutions in different countries.

Work package leader: MICF .

The aims of WP 7 – to test and evaluate how these objectives attained.

Task 4.3: Creation of virtual documents for research and teaching purposes (m6-m18)

4.3.1. Pilot implementation creating virtual documents for research and teaching purposes with existing technical resources – creation of sample virtual documents relating to the results of standardization (WP3) and publication of tools for free download; evaluated in respect of

a) Adaptability /Tools.

4.3.2. Approving of possibility to import documents into special collections in *Manuscriptorium*; evaluated in respect of:

b) Adaptability to digital objects,

c) Security of processing.

Hereafter at http://www.musicalia.lt/eta/wp4.php the respondent find three sets of questions to evaluate two main functionalities developed within WP 4 activities. The first functionality concerns the possibility for the end-user which enters *Manuscriptorium* to create personalized thematic collections and the second one allows the end-user to produce new virtual documents by merging different parts coming from diverse contents.

All the features are available in the Manuscriptorium Digital Library environment directly. The virtual documents have to be created using the M-Tool application and then later these can be sent to the Manuscriptorium Digital Library for publication/use. The respondents were advised to read the user guides: User Guide to the Personal Library features and User Guide for digital facsimile browsing. The questions are displayed at the evaluation web site [9] and added to this report in the Annex 1-c.

3.3.3. Personalization for Contributors – WP 5

Objectives

Development and implementation of the next generation of tools for structuring of existing metadata and newly created digitized documents and their implementation into Manuscriptorium structures, adjusted according to the results of WP3, accompanied by the use of large external data sets.

Providing on-line tools for verification, authentication and implementation of metadata and data into Manuscriptorium while respecting the specific characteristics of original data structures.

Allowing the accessibility of partners' data via Manuscriptorium without the need of changing their original presentation and use.
Work package leader: AIP.

The M-Tool can be accessed online**.**
Additional information **about evaluating M-Tool.**
The M-Tool User's Manual **can be accessed as a PDF.**

The aims of WP 7 – to test and evaluate how these objectives attained.

Task 5.1: On-line tools for structuring of existing metadata and data related to manuscripts (m3-m16)

5.1.1. Set of tools will be based on existing tool provided in the frame of Manuscriptorium; evaluate including on-line versions. Tools to be used by the content partners for structuring of existing data, are they well usable? Evaluated in respect of:

a) Adaptability/Tools,

b) Usability / Objects,

c) Usability / Tools.

The detailed questions are displayed at the evaluation web site [9] http://www.musicalia.lt/eta/wp5.php .

The second evaluation, denoted as WP 5 new, was split into two branches, corresponding to the means which content providers have used while submitting the content to repository: http://www.musicalia.lt/eta/wp51.php and http://www.musicalia.lt/eta/wp52.php, where the following aspects were evaluated:

a) INETROPERABILITY / Object,

b) ADAPTABILITY / Tools,

c) USABILITY / Processes,

d) USABILITY / Tools,

e) ADAPTABILITY / Object,

f) USABILITY / Object.

The questionnaires and the results obtained are added to this report in the Annex 1-d.

3.3.4. Multilingual and User Friendly Sophisticated Access – WP 6

Objectives

This work package aims at integrating a multilingual module via a user friendly sophisticated access: multilingual search application, multilingual forums, and multilingual ontology editor.

Based on SYSTRAN's machine translation technology, this module will provide also terminology extraction and machine translation customization tools for the construction and retrieval of personalized metadata within the aim to create new multilingual digital documents and multilingual ontologies in Czech, Polish, Spanish, Portuguese, German, Italian, English, French, Danish, Hungarian, Russian and Serbo-Croatian.

Work package leader: SYS

The aims of WP 7 – to test and to evaluate how these objectives attained.

Task 6.1: Multilingual access development (m0-m12)

6.1.1. Multilingual access via the API integration in the data retrieval interface associated or independent of a multilingual search. Evaluated in respect of:

a) Multilinguality in object level,

b) Multilinguality in processing.

6.1.2. Multilingual access dedicated translation interface where ENRICH expert users can fine-tune dynamically the machine translation tools thanks to adapted linguistic tools for terminology extraction and translation post-editing and customization. Evaluated in respect of:

c) Multilinguality / Tools,

d) Multilinguality /Repository.

For translation evaluation purposes SYSTRAN developed a translation evaluation interface which is accessible to all ENRICH users so as to evaluate the translation quality.

SYSTRAN Enterprise Server 6 enables users to submit feedback on translation quality, and thus raise issues and propose alternative translations. Once submitted, this feedback can be reviewed by a Dictionary Manager, who will thereafter be responsible for following its life cycle.

The Dictionary Manager determines the validity of submitted feedback. They can then quickly update the linguistic resources with proposed terminology entries and/or forward the feedback to SYSTRAN.

Following the submission, the feedback is entered into a database and an email is sent to the sender if the "Send me" notifications check-box was ticked at submission. The Search Feedback utility can be granted to users by a system administrator. Once in place, users access this tool via a Search Feedback command present in the left-hand Feedback menu. The Search Feedback page will display, offering a wide variety of search parameters.

The Feedback Search Results page will display, offering a list of feedback matches, as well as a number of tools that can be used to take action on the results selected.

A number of different actions can be performed on a feedback received, including:

- Exporting the selected feedback items
- Replay the translation of the selected feedback items
- Sending the feedback to SYSTRAN support
- Providing additional comments
- Changing the feedback status

All modifications occur following submittal. This enables WP 6 leaders to react effectively to all users' remarks.

The detailed questions are displayed at the evaluation web site [9] http://www.musicalia.lt/eta/wp6.php and added to this report in the Annex 1-e

3.3.5. Description of Evaluation Activities, the Outputs' Interpretation

- As soon as the indicated Tasks were accomplished, the partners were asked to evaluate the results using the rating from **0 to 4** to each of the questions (including (a) - (f) aspects where they are available) raised in the evaluation web site [9] for each of the packages: WP 3, WP 4, WP 5, and WP 6.
- All together the number items for evaluating the results in any WP 3 – WP 6 are in use, each of them can be classified as reflecting one or more of the main Criteria: Interoperability, Adaptability, Usability, Security and Multilinguality with their corresponding categories.
- The each item's score contribution range from **0 to 4** and the maximal possible score for every WP is calculated. When we prefer to have a result easily comparable with others, let us use the overall value of score obtained in WP as the ratio to maximum possible value or in percentages – ranging from **0** to **100**.
- The WP' or criterion' score falling in the interval means:
    - **0 – 25** that the result is low,
    - **26 – 75** it is rather good (satisfactory),
    - **76 – 100** it is very good.
- The visualization of the achieved results in respect of the main principles of quality such as Interoperability, Adaptability, Usability and their categories are done by adding these results to the evaluation of other WPs. These results are provided in the next chapter.
- **The results – evaluated and validated as it was foreseen in the WP 7:**
    - **WP3 developed Migration Tools;**
    - **WP4 personalized tools for creation of virtual documents by researchers;**
    - **WP 5 basic conditions for sharing of large data sets;**
    - **WP6 Personalized Translation Interface.**
- The permanent monitoring of the evaluation process were performed, published at the evaluation site in statistical pages http://www.musicalia.lt/eta/stats.php and periodically sent by e-mails to project partners participating in the evaluation activities.

3.4. Results of Evaluation, Testing and Validation for ENRICH WPs and Quality Criteria

Summary of Evaluation Results

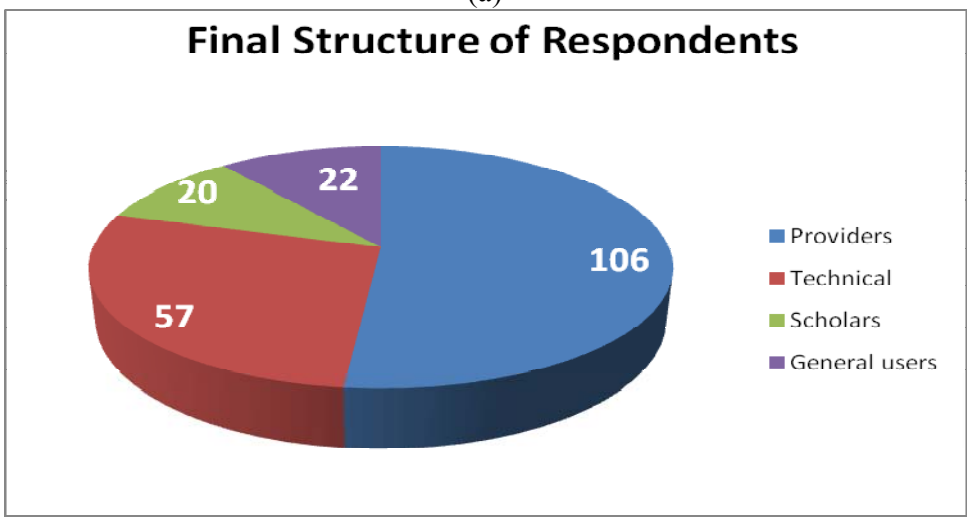
In total **205** respondents were involved into process of evaluation by filling the evaluation forms online during the evaluation process from April 2009 to November 2009. Each respondent had to express own opinion on the specified result, assigning a score – an integer ranging from 0 (poor or not available) to 4 (excellent). The double average values over respondents and questions evaluated have to ensure more stable estimators of quality. The structure of respondents profile and their general opinions on results achieved during 24 months of project work are illustrated in the Fig.5 – 7. The results of evaluation were analyzed in many aspects and are presented in this report: illustrated by Fig.5 – 26 with the corresponding comments to each picture and the final conclusions at the end.
First of all, the results were considered across the different users' groups in order to investigate how their needs were satisfied in general and in each work package, the diagrams are shown in Fig. 5 – 20. Secondly, the average scores, assigned to the separate questions on results achieved in WPs by all target groups of respondents, were recalculated as estimates of Categories and Quality Criteria over investigated WPs, those results are shown in the Fig. 21 – 22. In order to have statistically reliable inferences, the 0,95 confidence intervals were fitted to the estimators of quality as they were considered in all aspects: WPs, Criteria, Categories and opinions of target groups. Those results are displayed in the Fig. 23 – 26.

It is shown that with the 0,95 confidence level, based on 205 respondents answers, we can confirm that the created Processes, Tools and Objects are the leading properties achieved among the ENRICH results. Similarly, the Interoperability and Adaptability are the best (with the 0,95 confidence level) in ENRICH results when compared with other quality Criteria such as Multilinguality or Usability. The numerical values of estimated quality aspects are shown in created diagrams using two ways: the average scores ranging from 0 to 4 (Fig. 5 – 23) and their percentages (Fig. 23 – 26) with the lower and the upper confidence limits indicated.

3.4.1. The Results across the Different Target Users Groups

The total number of the respondents involved into evaluation of ENRICH results and testing activities was **205.** The four target users groups were considered: content providers-information managers (106), technical personnel-supporting staff (57), and scholars – researchers in historical documents, students (20), and the general or the end-users having general interests (22). The structure of respondents corresponds well to the aims of the project which is oriented more to the experts in the area than to the users of a general interest. Regretfully, the activity of scholars during whole testing period was rather low; it is more comparable to general users than to experts. But certainly it reflects a real structure of users because the ratio of target groups was rather stable during the evaluation process as seen from the Fig. 5, the cases (a) and (b).

(a)



(b)

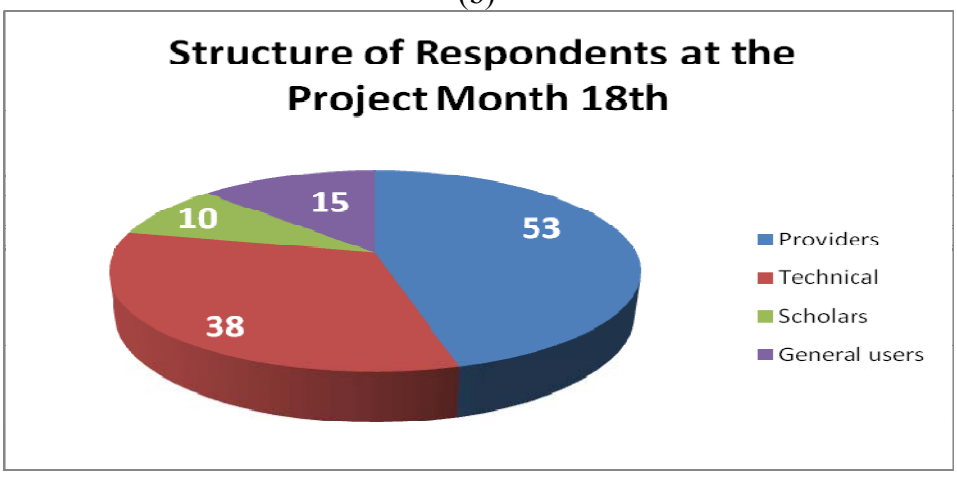


Fig.5. The proportions of different users evaluating results WP 3 – WP 6 were almost stable: (a) The distribution of **205** respondents over the four target users groups: content providers – information managers, technical personnel – supporting staff, scholars – researchers in historical documents, students and the general or the end-users having general interests; (b) The distribution of **116** respondents at the 18th month of project work.

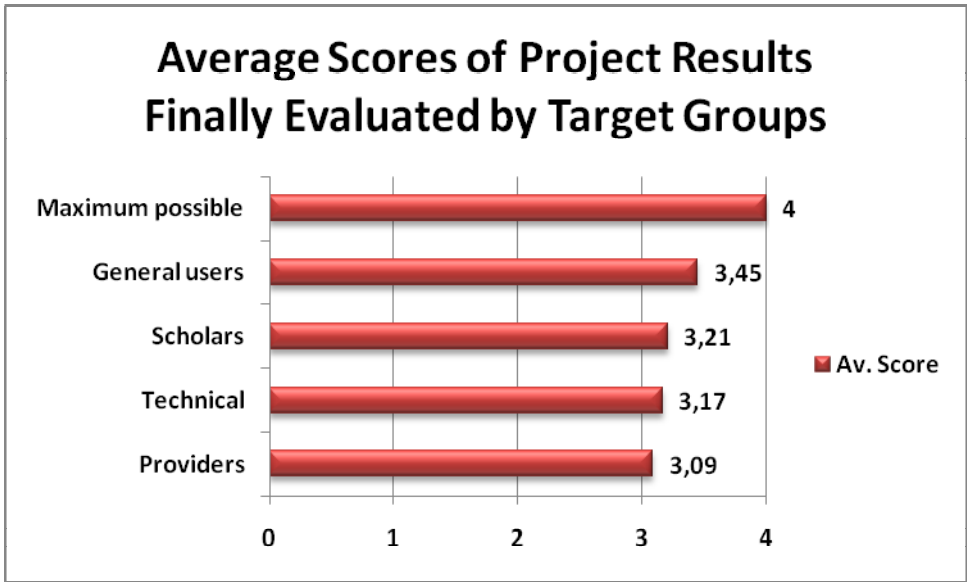


Fig.6. Project results in WP3, WP4, WP5, WP5 new, WP6 as evaluated by target users groups, compared to the maximum possible score 4.

The summary results shown at the Fig.6 demonstrate rather similar opinion of scholars, technical staff and content providers while general users seem to be more satisfied with ENRICH project results than the experts. The results are rather similar to those derived at the month 18th of ENRICH work (6 months to the project end). In the 3.4.4 section of this report we will consider the confidence limits for each group of users in order to make a statistically correct conclusion. The data from the questionnaires filled on-line by the project partners are shown in the Fig.7. What conclusions can be made about a quality from such data? It would be difficult to judge from such raw data – detailed statistical analysis and investigation were made and described in this Evaluation Report, enabling to make statistically correct inferences for a quality matters evaluated.

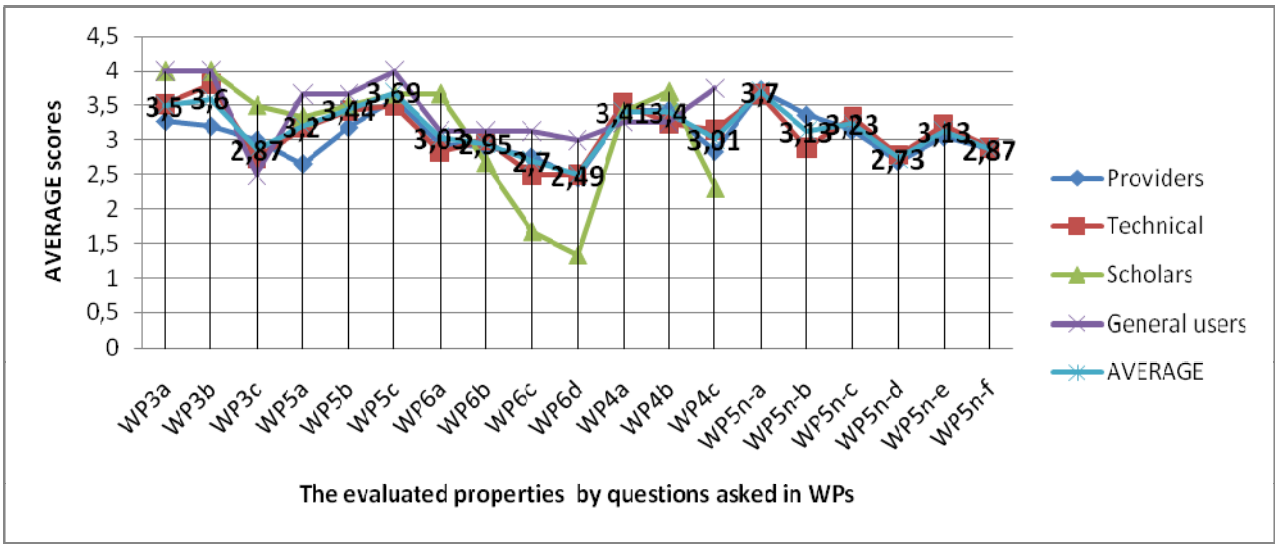


Fig.7. The investigated properties (19 questions asked) concerning the quality in WP , WP4, WP5, WP5 new and WP6, evaluated by target users groups. The average values over target groups are shown numerically on the blue line.

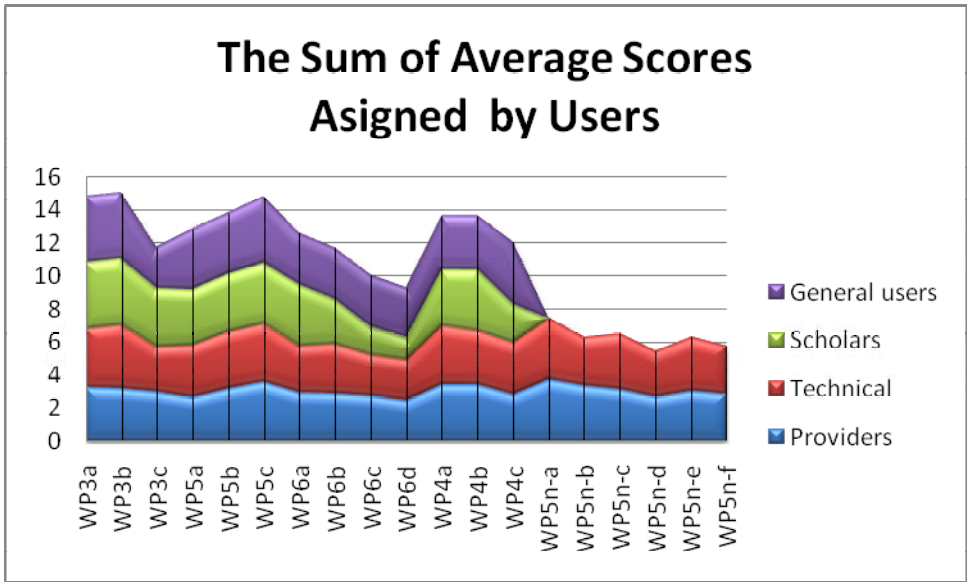


Fig.8. The sum of scores of 19 sub criteria, each having 4 questions asked, as they were evaluated by 205 respondents. The total number of questions answered is 76.

The sums of scores in Fig.8. is the highest at **WP 3b = 15,02** and **WP 5c = 14,76**. The spread of the opinions among the users groups is rather equal in the WP 3, WP 4 and WP 5 questions but have much larger variations in the WP 6 quality evaluation. The smallest summary value is for **WP 6d = 9,28** it contains also the smallest question score 1,33 assigned by scholars. The last evaluation concerning questions from WP5n-a to WP5n-f was too difficult for general users and was done only by experts' groups.

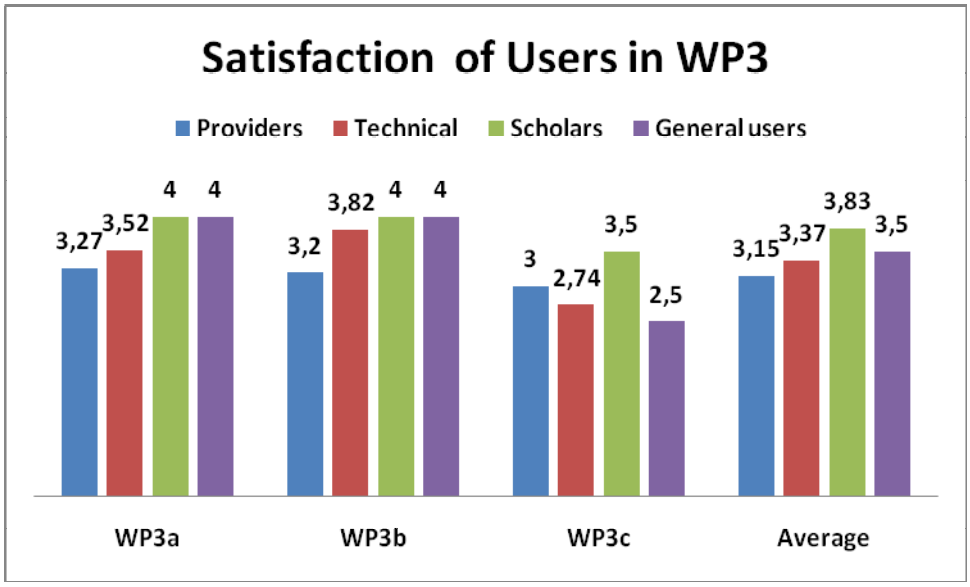


Fig.9. The averages of the numerical values, assigned to each of the three questions, reflecting the quality of WP 3 results, evaluated by different users groups.

The average of WP 3 (located at the right of the diagram, Fig. 9) shows the spread of opinions on quality in WP 3 by target users. The total average of WP 3 is equal to **3,32.** The questions in WP 3 were rather technical and certainly difficult to access for general users and sometime for scholars – researchers in the historical documentary heritage. Investigating more

thoroughly the results in Fig. 9 we see that the four extreme values equal to 4 are allocated at WP3-a and WP3-b by scholars and general users and they can affect significantly the average of WP 3. Let us exclude the extreme values and apply the stratified sampling for evaluating WP 3 results only by real experts: content providers, information managers (15 respondents in WP 3) and technical personnel, supporting staff (17 respondents in WP 3). Happily, they were in majority of this sample (compared to 2 scholars and 4 general users only).

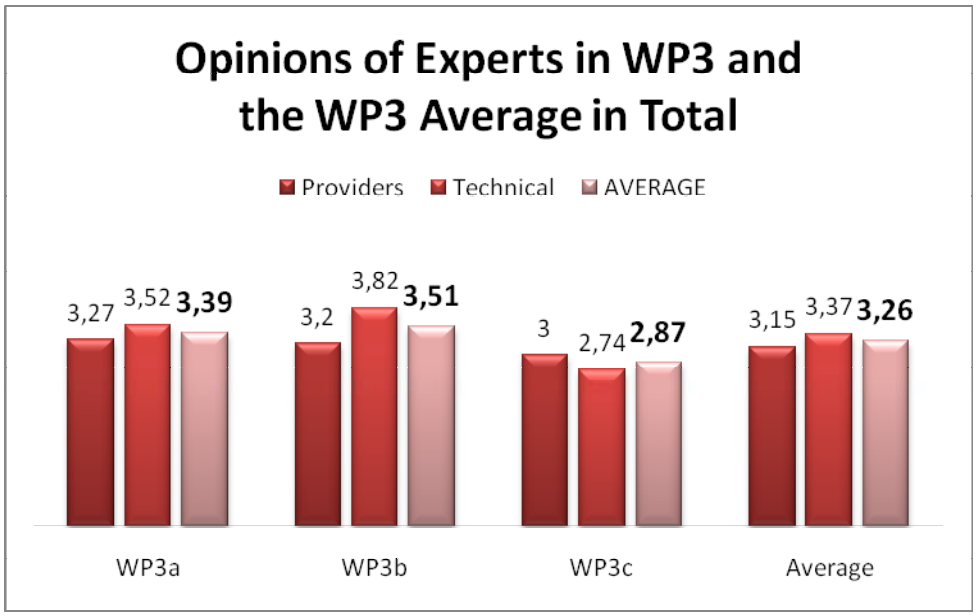


Fig.10. The averages of the numerical values, assigned to each of the three questions, reflecting the quality of WP 3 results, evaluated only by expert users groups.

The average at the right of the diagram, Fig.10, shows the total averages assigned by content providers and technical staff and the total average of WP3 given by experts is equal to **3,26**.

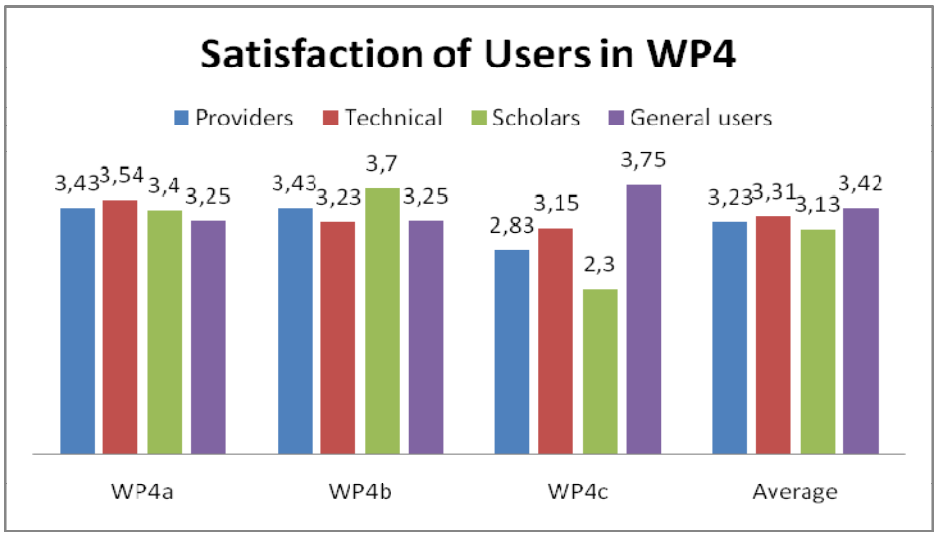


Fig.11. The averages of the numerical values, assigned to each of the three questions, reflecting the quality of WP 4 results, evaluated by target users groups.

The average calculated across the WP 4 questions shows the total evaluation of a quality in WP 4. The average of WP 4 in total is equal to **3,27**.

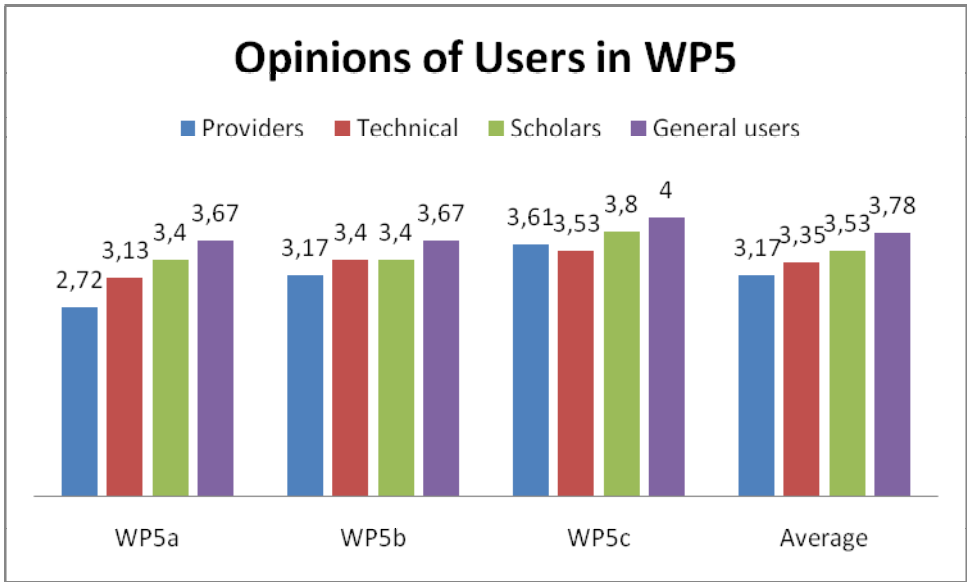


Fig.12. The averages of the numerical values, assigned to each of the three questions, reflecting the quality of WP 5 results, evaluated by target users groups, 49 respondents.

The average calculated across the WP 5 questions in Fig.12 shows the total evaluation of a quality in WP 5. The average of WP 5 in total was equal to **3,32** at the first evaluation.

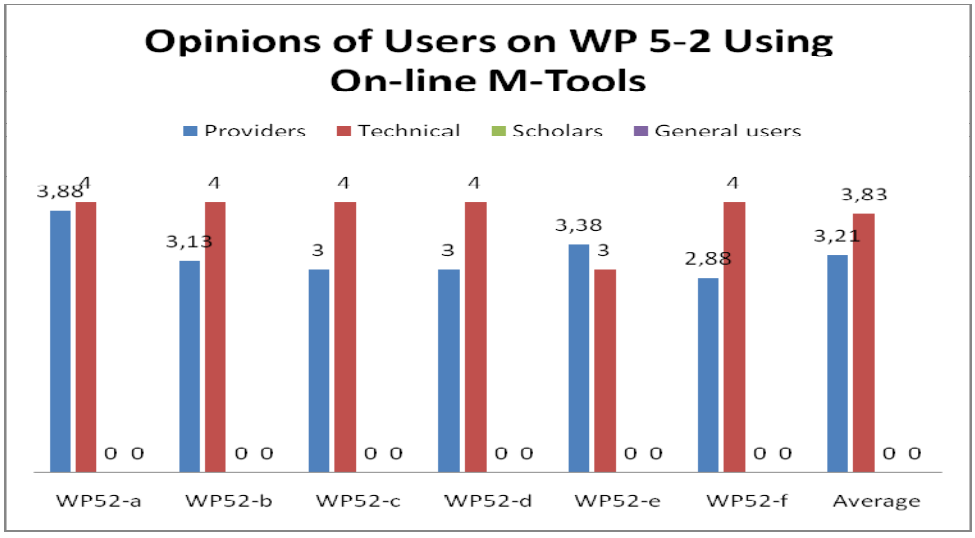


Fig.13. The evaluated WP5n quality (using on-line M-Tool) in the second evaluation.

The questions in WP 5 new evaluation were so specific that users of general interest and scholars were not able to answer them and the number of target groups naturally becomes smaller and equal to two expert groups – content providers and technical staff (instead of previously stated four groups).

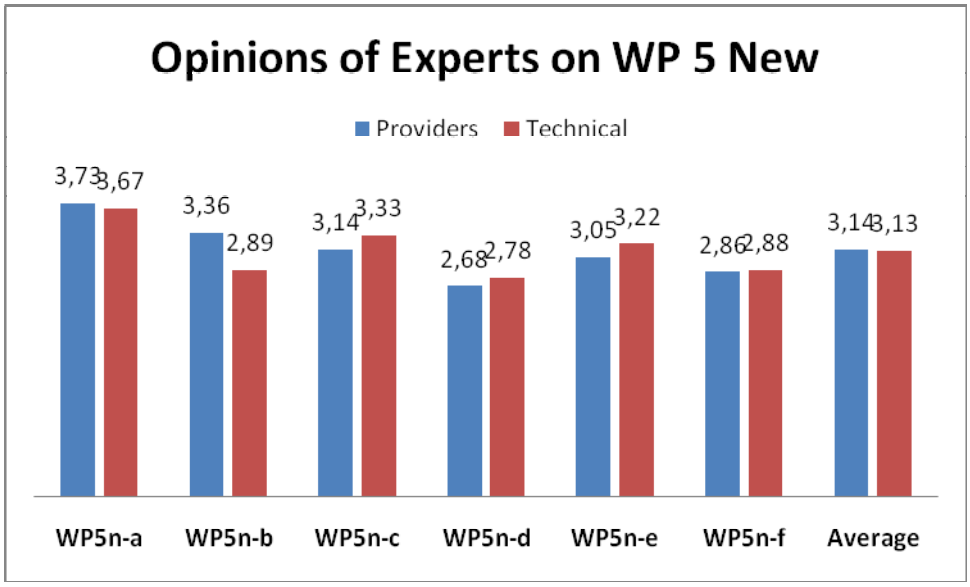


Fig.14. The averages of the numerical values, assigned to each of the six questions, reflecting the quality of WP 5 new evaluation results, evaluated by the expert users groups.

The Fig. 14 shows the results provided by 31 respondent – content providers and technical staff participating in both versions of evaluation held in the second evaluation. They have almost the same opinion, as an average at the right of diagram shows evidently. The new evaluation resulted in total to average value **3,135**.

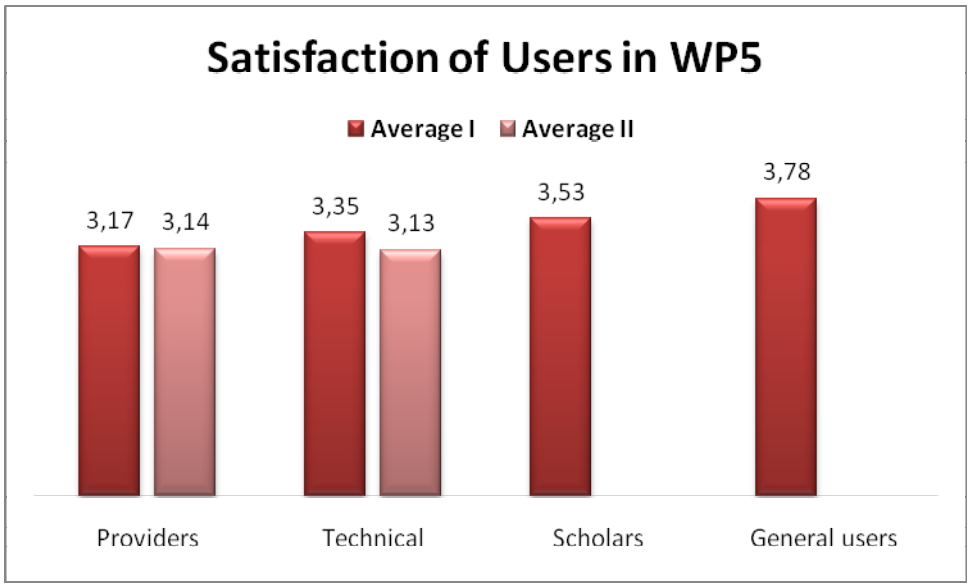


Fig.15. The averages of the numerical values, assigned by users to facilities created in WP 5 during the first (I) evaluation and the second (II). Users of general interest have been very optimistic during the first evaluation (resulting in 3,78 score from the maximum possible 4).

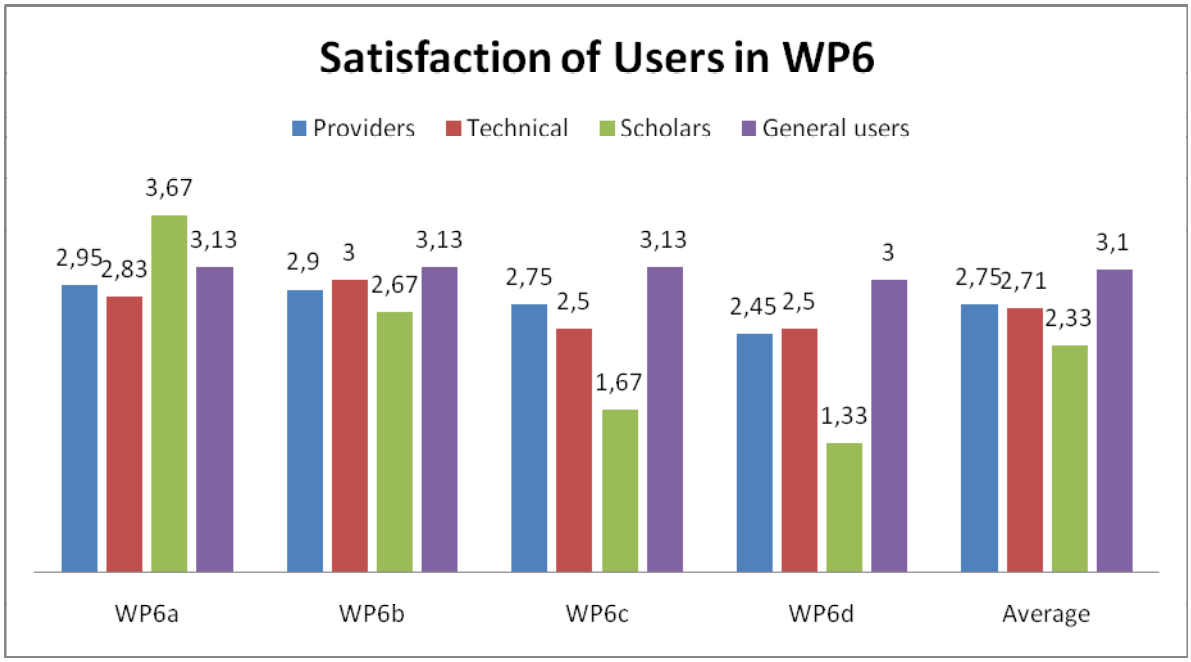


Fig.16. The averages of the numerical values, assigned to each of the four questions, reflecting the quality of WP 6 results, evaluated by target users groups, 37 respondents.

The average values located at the right of the diagram in Fig. 16 provides the total evaluation of a quality in WP 6 by different users. The average of WP 6 in total is **2,66.** This is the lowest average obtained from all evaluations performed in WPs.

3.4.2. The Scores in Work Packages Assigned by All Users

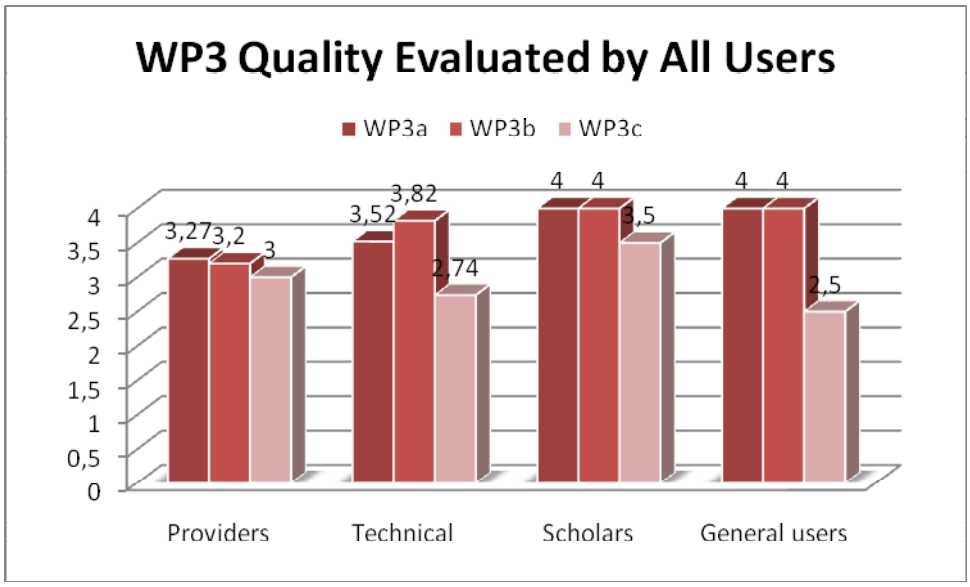


Fig.17. The average scores of 3 questions on a quality in WP 3 evaluated by full sample (38 respondents) of users. The four extreme values equal to 4, assigned by scholars and general users, located in the WP3-a and the WP3-b.

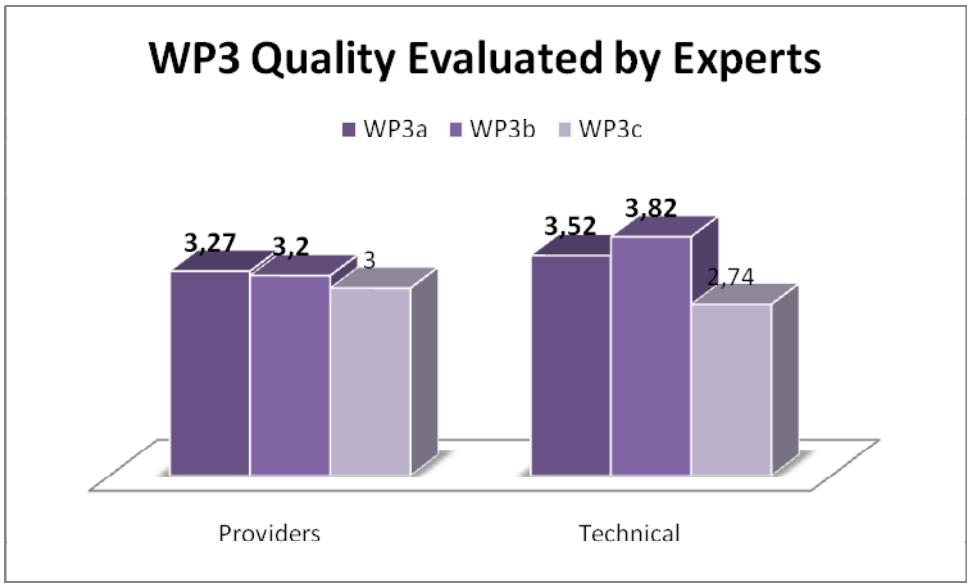


Fig.18. The average scores of 3 questions on a quality in WP 3 evaluated by stratified sample including only the expert users. The total number of respondents there is 32. The extreme values, affecting the final average, were excluded. By experts evaluation the average of WP 3 in total is **3,26** (let us compare it to the previous result equal **3,32**).

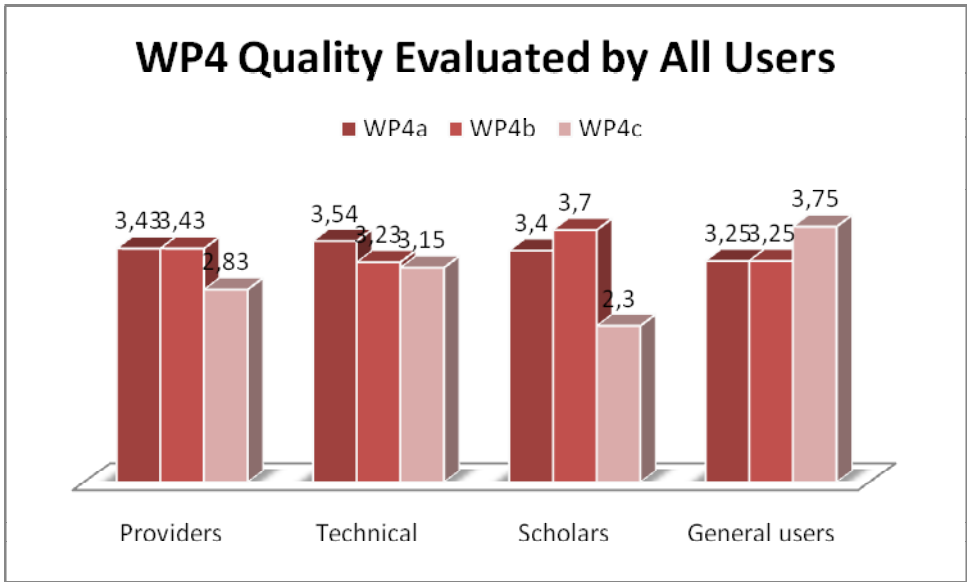


Fig.16. The average scores of 3 questions on a quality in WP 4 evaluated by a sample (50 respondents) of users. The average of WP 4 in total is equal to **3,27**.

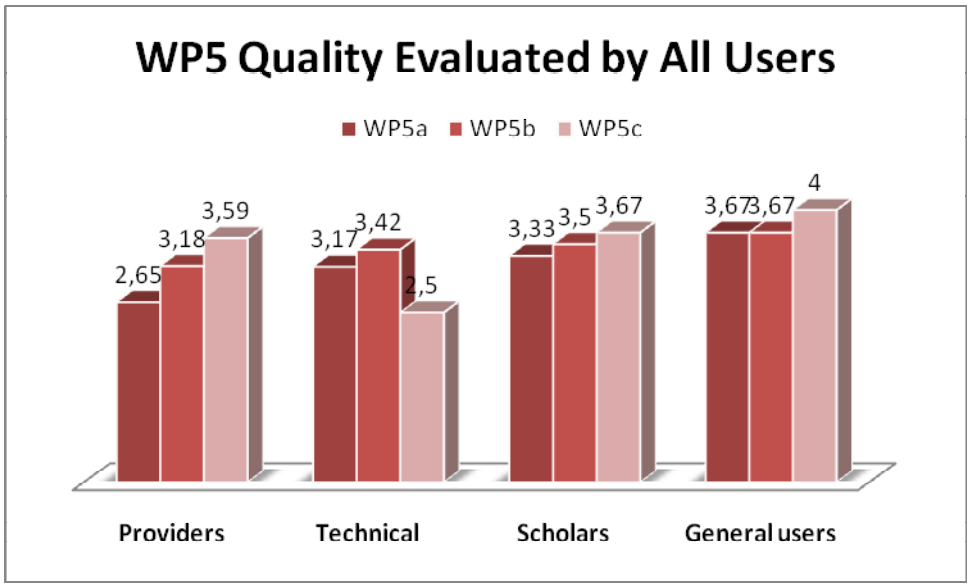


Fig.17. The average scores of 3 questions on a quality in WP 5 evaluated during the first evaluation (by 49 respondents) by all users of WP 5. The general users once more assigned rather high scores to all questions.

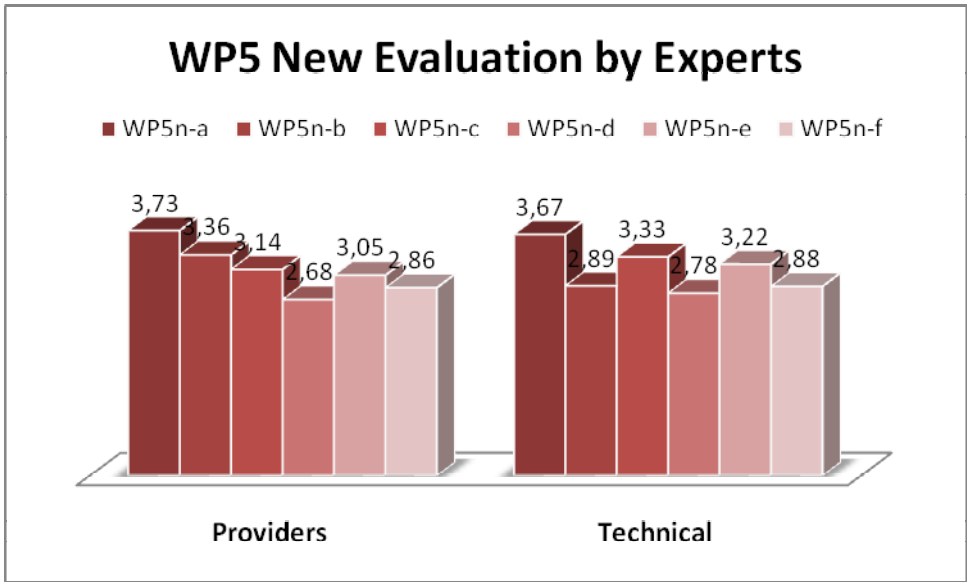


Fig.18. The average scores of 6 questions on a quality in WP 5 during the second evaluation (31 respondents) done by the expert users of WP 5n.

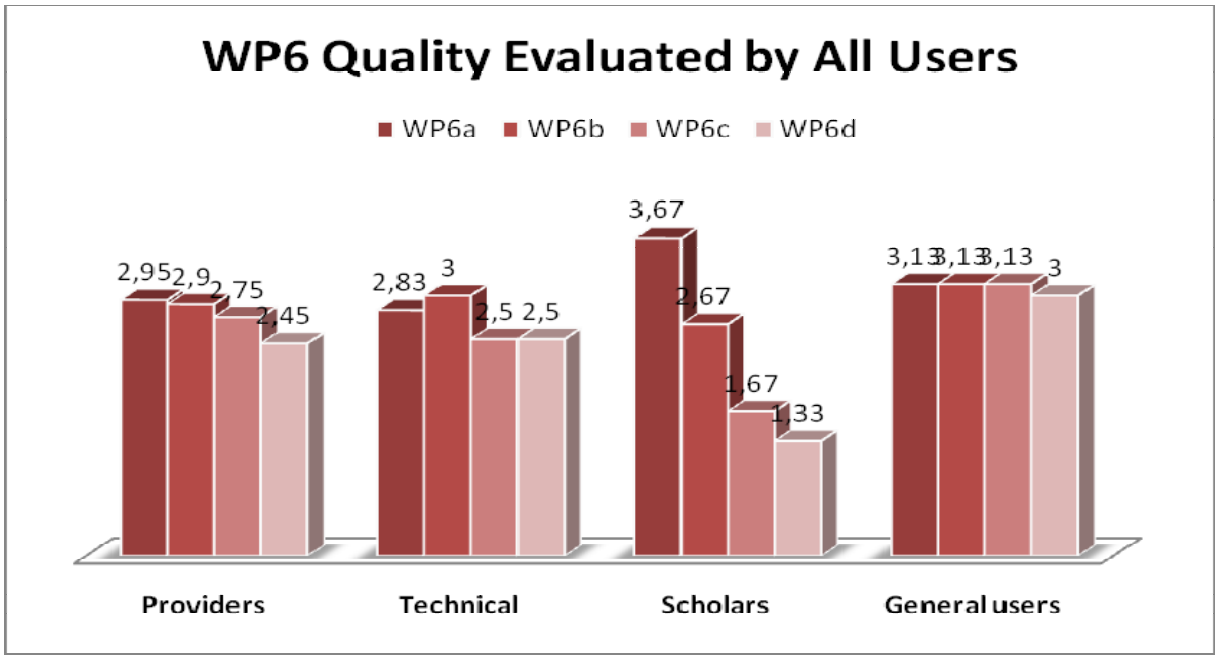


Fig.19. The average scores of 4 questions on a quality in WP 6 as were evaluated by all users (37 respondents). The scholars have the most spread opinions in contrast to the general users having an opinion almost uniformly good.

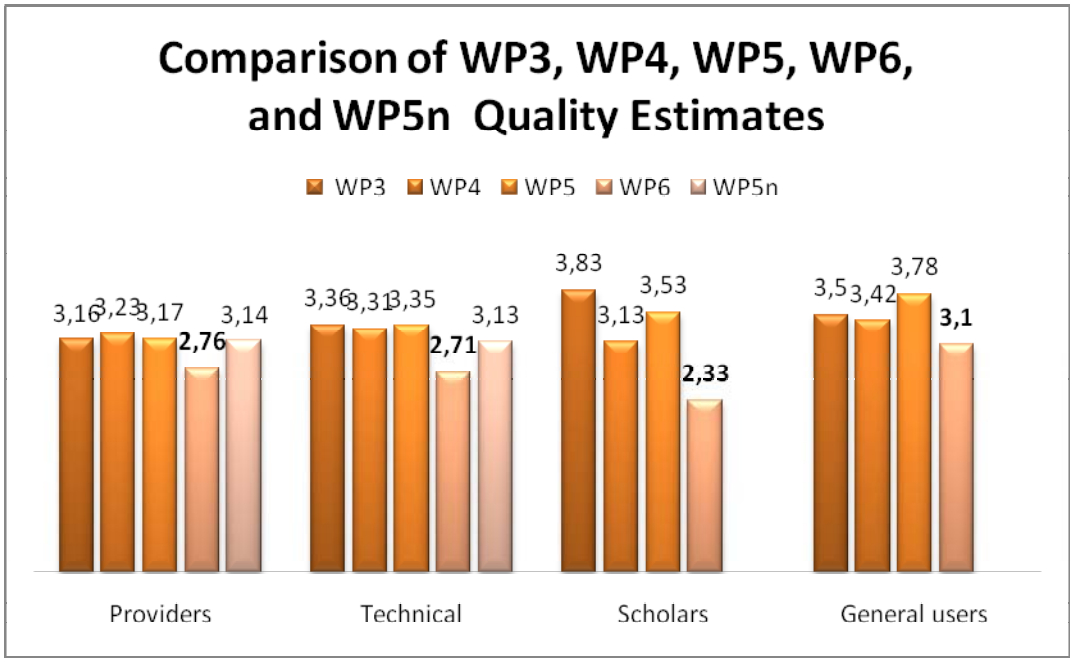


Fig.20. The average total scores, reflecting a quality in WP 3, WP 4, WP 5, WP 5 new, and WP 6, evaluated by different users in the pooled sample of 205 respondents.

It seems that the quality of results achieved in the WP 3 and WP5 were evaluated higher than in WP 6 – that is an integrated opinion of 205 respondents involved into the final evaluation and illustrated in the Fig.20. This conclusion is confirmed later by fitted confidence intervals to estimates and shown in the Fig.23.

Let us compare the total averages given by various users to WP 3, WP 4, WP 5 and WP 6. Those quantities are 3,32 (or 3,26 evaluated only by experts) 3,27, 3,32, 3,135 and 2,66, respectively. Are those estimates significantly different? The problem of statistical inference is as follows: testing a hypothesis that WP 3 results are better than those of WP 6 will be addressed in the 3.4.4 section of this report and demonstrated in the Fig. 23.

3.4.3. The Scores in Categories and Main Criteria

Now let us derive the average scores across the four categories (digital objects, tools developed in ENRICH project, processing, and a repository as a whole) and five Main Criteria in quality from all collected data from 205 respondents participating in the testing and evaluating activities and reflecting their opinions during several sessions of evaluation. The evaluations of results, achieved in the project ENRICH Work Packages, namely the WP 3 – WP 6, started in April 2009 and were finished in November 2009. The WP 5 was evaluated May – June, and repeatedly in October - November 2009, WP 4 – September 2009, WP5 new finished to the end of November 2009.

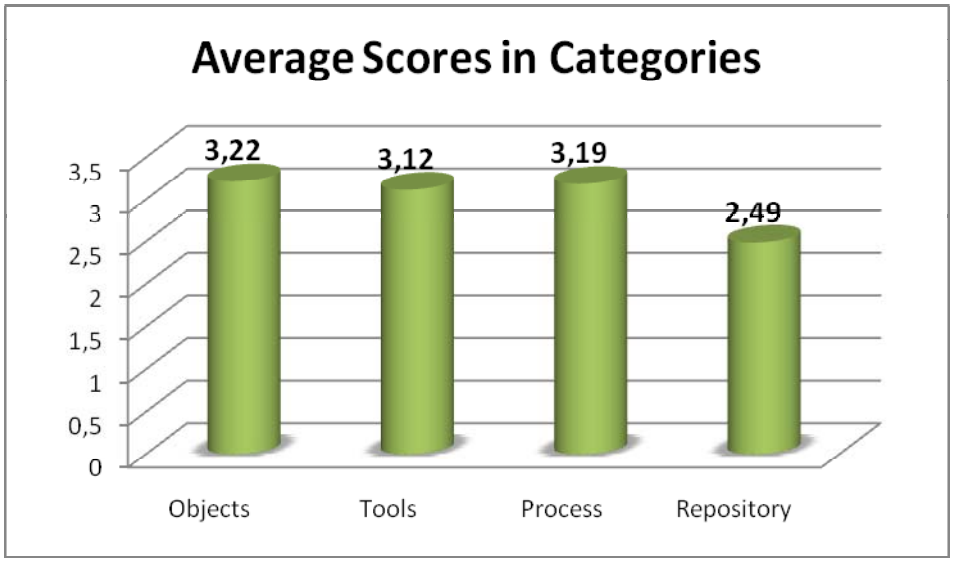


Fig.21. The average scores of the four categories as the components of a quality in ENRICH project results achieved in WP 3, WP 4, WP 5, WP 5n, and WP 6 extracted from a pooled sample of 205 respondents.

The maximum possible score for the results shown in the Fig.21 is 4. Visually the weakest point is in the category *Repository*. Is it true? – it will be checked in the 3.4.4 section of this report by constructing the confidence intervals for each category.

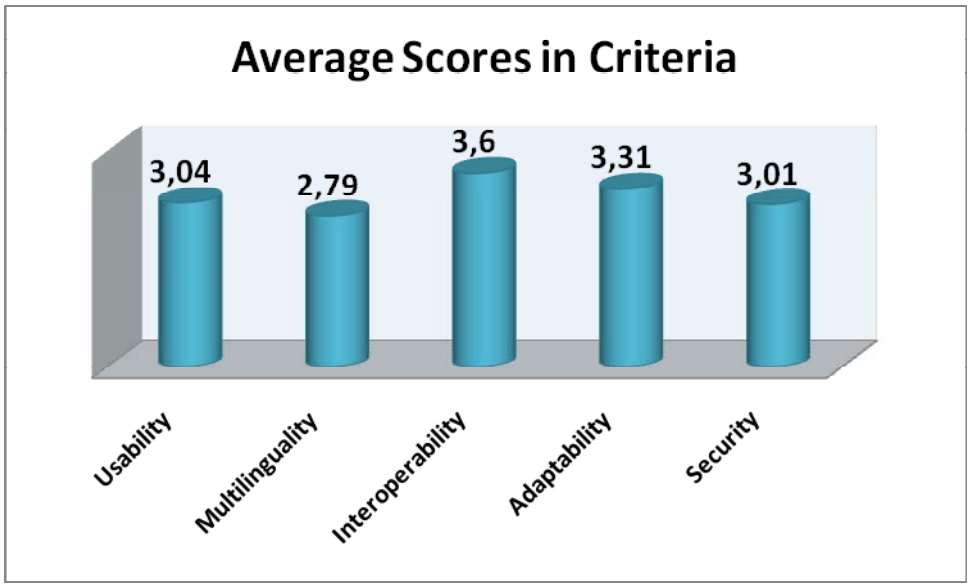


Fig.22. The average scores of the five Main Criteria reflecting the quality of ENRICH project results, achieved in WP 3, WP 4, WP 5, WP 5n, and WP 6 extracted from a total sample of 205 respondents obtained during several evaluation sessions.

The Interoperability and Adaptability in the Fig.22 seems to be estimated better than Usability and Multilinguality or Security. Are those differences significant? The question of correct comparison of available evaluation results in WP, Categories and Criteria will be answered in the following section.

3.4.4. Statistical Inference on Derived Results – Confidence Limits of Estimators

In order to test correctly the statistical hypothesis, say that the WP 3 results are better than those of WP 6, let us fix the standard significance value 0,05 corresponding to 0,95 confidence level (and the critical value 1,96 in the normal approximation of the statistic used). Then the following approximate 0,95 confidence interval, investigated specially [6, 7] for the case what we come across here is fitted as:

$$p^* \pm 1{,}96\,[p^*\,(1-p^*)\,/\,4nk]^{1/2}.$$

Here p^* is an average quality estimator divided by it's maximum possible value, n – number of respondents, k – the number of questions (sub criteria) used for evaluation. More details are given in the section 3.2.2. Applying this formula we have the confidence intervals for WPs investigated, summarized in the Table 2.

Table 2. Total Averages Evaluated in WPs and Their Confidence Limits.

	Lower confidence limit	Total average evaluated	Upper confidence limit
WP 3 (all 38 respondents)	3,29	3,32	3,35
WP 3 (32 expert respondents)	3,22	3,26	3,30
WP 4	3,24	3,27	3,30
WP 5	3,29	3,32	3,35
WP 5n	3,11	3,135	3,16
WP 6	2,62	2,66	2,70

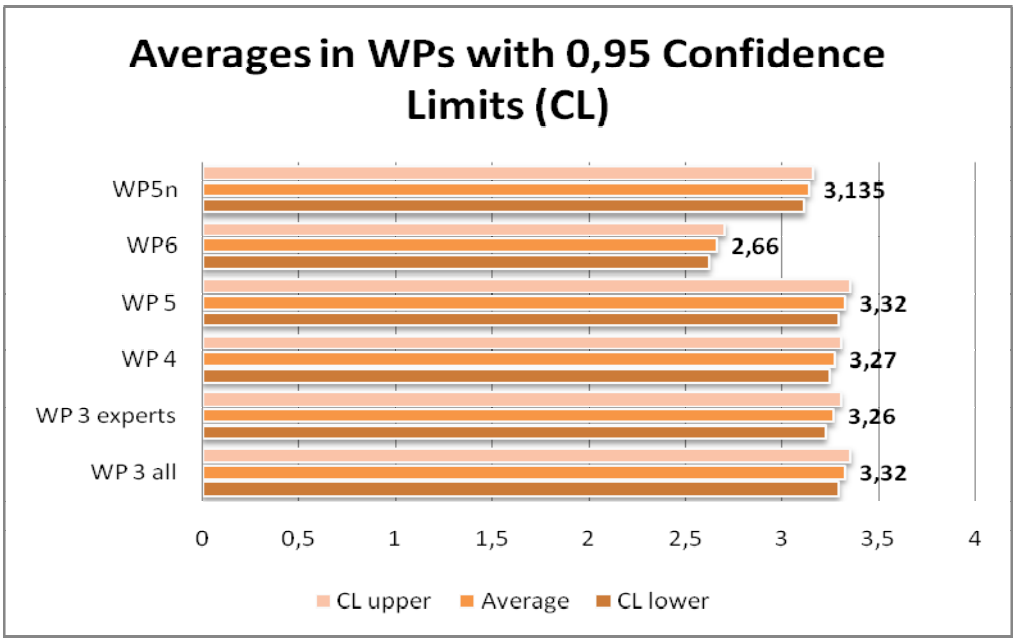


Fig.23. The averages of a quality in the ENRICH project results, achieved in WP 6, WP 5, WP5n, WP 4, and WP 3 as evaluated by all users (by experts only in WP 3 and WP 5n).

The conclusion what follows from Table 2 and it is seen evidently from the Fig. 23, where the lower and upper 0,95 confidence limits are displayed, is the following. Because the confidence intervals for WP 3 results (in both cases: all respondents and only experts) and WP4, WP5 are overlapping we can conclude that there is no significant difference among them. WP 5n results are similar to the mentioned above. But evidently a different conclusion follows when considering a difference in a quality of the WP 6 and other WPs – it is significant and with the probability equal to 0,95 we can confirm that results in WP 3 (or WP 4, WP 5) are better estimated than in WP 6. The approximate upper and lower confidence limits (CL) are derived for each case and displayed in Fig. 23. The results are related to those in Fig.20 and show evidently that the significant difference in a lower quality what can be concluded is in the WP 6. Other results of a quality are comparable to each other with the confidence level 0,95.

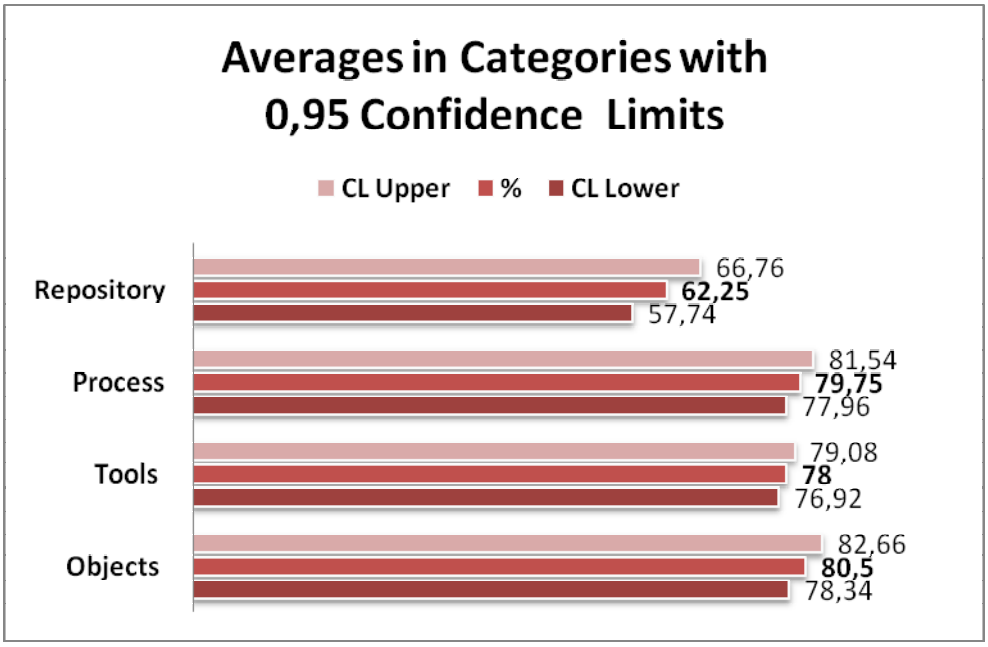


Fig.24. The percentages of the four categories as the components of a quality in the ENRICH project results, achieved in WP 3, WP 4, WP 5, WP 5n and WP 6 as they were evaluated by 205 respondents. The approximate upper and lower confidence limits (CL) are derived for each category considered.

Results displayed in the Fig. 24 show that we can conclude with the 0,95 confidence level that the developed Processing, Tools and Objects are equally well evaluated as ENRICH results, especially when compared with using Repository as a whole. Let us notice that the Repository evaluation is also rather good. Looking in general, all these aspects were evaluated well, the points assigned to those categories (from the maximum possible 100 points) all are **good**. According to the Methodology for Evaluation, shortly described in the 3.3.5 section, the score falling in the interval 0 – 25 means that the result is low, 26 – 75 it is rather good (satisfactory), 76 – 100 it is very good. Therefore the Tools, Objects and

Processing received **very good** evaluation from the ENRICH partners and other related institutions, while the whole Repository – only satisfactory.

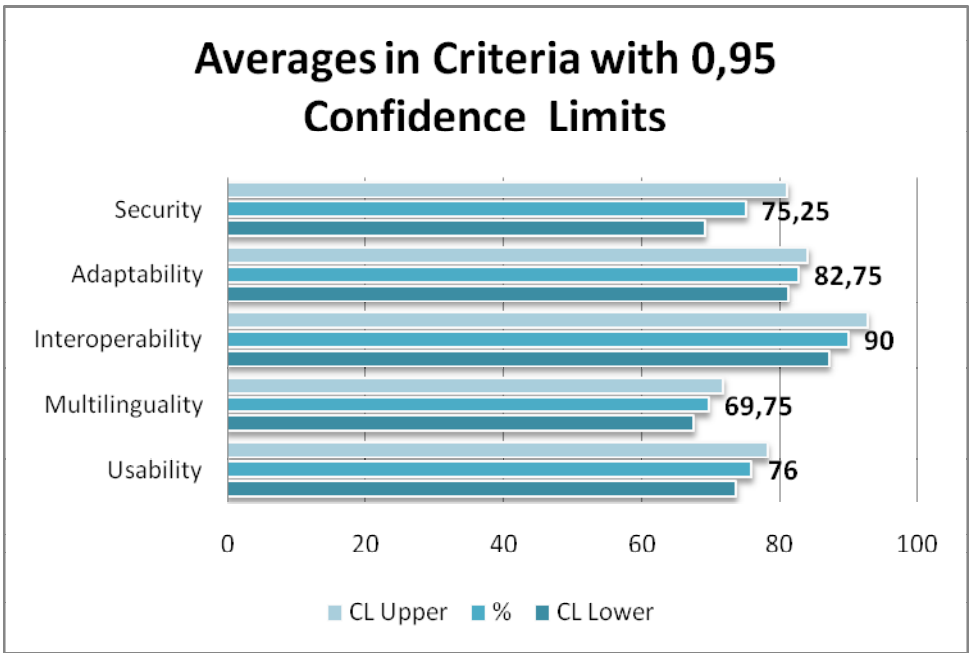


Fig.25. The percentages of the five Main Criteria reflecting a quality of the ENRICH project results, achieved in WP 3, WP 4, WP 5, WP 5 n and WP 6 as they were evaluated by 205 respondents during several evaluation sessions.

The approximate upper and lower confidence limits (CL) derived for each Criterion. Interoperability and Adaptability received very high scores and those are different from the evaluated Multilinguality or Usability properties with the 0,95 confidence level.
Results displayed in the Fig.25 show that we can conclude with the 0,95 confidence level that the Interoperability and Adaptability are the best ENRICH results when compared with other Criteria such as Multilinguality, Security or Usability. Looking in general, all involved Criteria have got the very high estimates, even the lowest result 69,75 assigned to Multilinguality (from the maximum possible 100 points) is **rather good**. The Interoperability, Adaptability and Usability received **very good** evaluation or the excellent mark from the ENRICH partners and users from other related institutions.

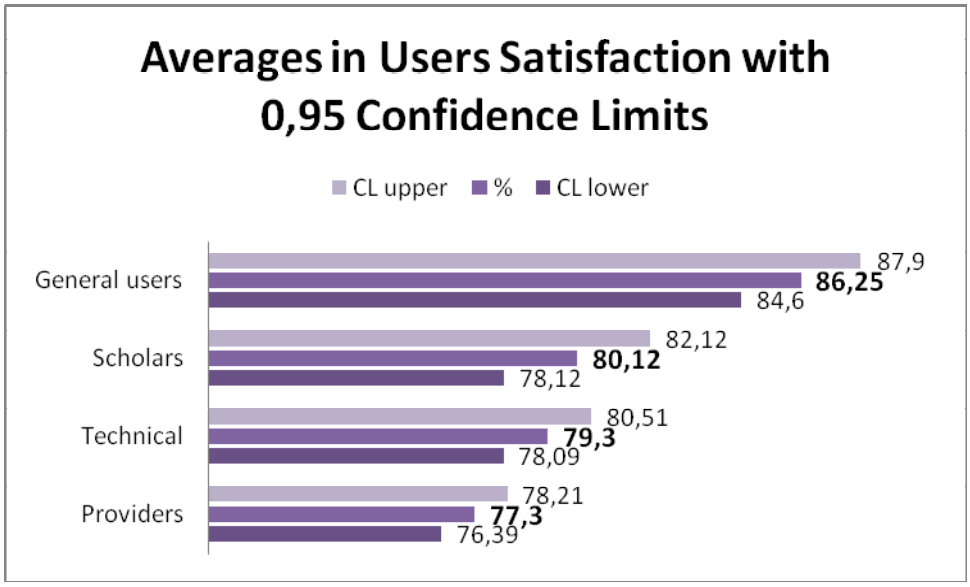


Fig.26. The percentages of satisfaction of target groups on a quality of the ENRICH project results, achieved in WP3, WP 4, WP 5, WP5n and WP 6 as they were evaluated by 205 respondents during several evaluation sessions.

The approximate upper and lower confidence limits derived for each group of users show that there are no significant difference among experts' users opinions – almost all intervals are overlapping. The real difference can be stated only between opinions of experts groups and the general users. The general users are more satisfied with the ENRICH project results than the experts – content providers, technical staff and scholars. The tendency that estimates of general users' are higher than other users groups were noticed in all aspects of investigation performed, but a real reason of this effect is unclear, one can only guess (but this lies outside our Evaluation Report). We can confirm only, that during the process of evaluation, at first stages it was not possible to make similar conclusions on target groups of users until sufficient sample sizes in the separate groups were not attained at the final stage of evaluation.

Finally we have the sample size equal to 205 respondents containing: content providers-information managers (106), technical personnel-supporting staff (57), scholars – researchers in historical documents, students (20), and the general or the end-users having general interests (22). This enables us to make a statistically reliable inference that general users are more satisfied with a quality in digital repository than more qualified expert users.

3.4.5. Conclusions and Comments of Obtained Results

1. The measures evaluating a quality of results achieved in the ENRICH project have been created and reflects a satisfaction of users in target groups using the results developed in project: digital objects, tools, processing and usability of whole repository as well as the main principles of quality: Interoperability, Adaptability, Usability, Security, Multilinguality.

2. Numerical evaluation results are comparable to each other and evidently show the weak and the strong points in the facilities developed or other digital repository aspects.

3. The summarized results reflecting the work in the project during 24 months, based on the 205 respondents opinions in total sample, enables to make statistically correct inferences with the 0,95 probability. They are the following:

- Interoperability and Adaptability received very high scores and those are significantly different from the evaluated Multilinguality, Usability and Security properties (Fig.25 and the *Table II* below).
- Processes, Objects and Tools were evaluated better than a whole Repository (Fig.24 and the *Table III*)
- The quality of the WP 6 results concerning multilingual access was evaluated to be significantly lower than other work packages: WP 3, WP 4, WP 5, WP 5n (Fig.23 and the *Table I*).
- All results are evaluated rather highly: the most optimistic were the general users, less – the experts (Fig.6). Statistically their opinions are different with the 0,95 probability (Fig.26 and the *Table IV*).

4. The fact that general users almost always were more optimistic while evaluating a quality in various aspects can be explained that specific, rather technical testing-validating tasks were hardly accessible to non-professional users – their estimates are too optimistic and could be not very reliable, but a validity of results on the differences among those target groups are proved statistically. The evaluations of experts are rather similar in all considered aspects and demonstrate their rather high satisfaction by functionalities and properties developed in ENRICH but they are less optimistic than general users' evaluations.

First of all the tasks formulated and performed in each work package (WP) were evaluated separately by asking users in the project partners institutions how they estimate the specified results achieved in that WP applying the scale from 0 to 4. The double average (over respondents and questions asked) was taken as an estimate of a quality. Those results of quality evaluation are summarized in the *Table I*.

Table I. ***WORK PACKAGES*** *– Estimated Scores and Their Percentages of the Maximum Possible Values*

Work Package	Estimate	% of 4
WP 3 – Standardization of shared metadata (38 respondents, 3 x 4 = 12 questions)	**3,32**	**84,25**
WP 4 – User personalization, creation of virtual documents (50 respondents, 3 x 4 = 12 questions)	**3,27**	**81,75**
WP 5 – Personalization for Contributors (49 respondents, 3 x 4 = 12 questions)	**3,32**	**84,27**
WP 5n – Aggregation of documents from each content provider (31 respondents, 6 x 4 = 24 questions)	**3,135**	**78,38**
WP 6 – Multilingual and user friendly access (37 respondents, 4 x 4 = 16 questions)	**2,66**	**66,5**

Those evaluations are based on the selected Tasks in WPs, namely: T 3.1, T 4.3, T 5.1, T 5.3, T 5.4, T 6.1, T 6.2 and, probably, reflects only a part of a work done in the separate WP. Therefore we do not focus on those estimates only. The main idea is to extract from those estimates the information related to the Main Quality Criteria and Categories. The relationships of WP 3 – WP 6 items and Quality Criteria were established in advance, in the Methodology D-7.1, and the estimates of Criteria and Categories derived from the pooled sample from several evaluation actions performed in the framework of the WP 7. Recalculated estimates, displayed in the Fig. 24, show that we can conclude with the 0,95 confidence level that the developed **Processes, Tools** and **Objects** are equally good evaluated as the ENRICH results, especially when compared with facilities in Repository as a whole. But the Repository evaluation is also rather good. Looking in general, all these aspects were evaluated well, the points assigned to those categories (from the maximum possible 100 points if we use percentages) all are **good**.

Similarly, results displayed in the Fig.25, show that we can conclude with the 0,95 confidence level that the **Interoperability** and **Adaptability** are the best ENRICH results when compared with other Criteria such as Multilinguality, Security or Usability. Looking in general, all involved Criteria have got very high estimates, even the lowest result 69,75 assigned to Multilinguality (from the maximum possible 100 points) is rather good. The Interoperability, Adaptability and Usability received **very good** evaluation or the excellent mark from the ENRICH partners and users from other related institutions. Let us remember that according to the Methodology for Evaluation (D-7.1) the score falling in the interval 0 – 25 means that the result is low, 26 – 75 it is rather good (satisfactory), 76 – 100 it is very good.

Considering a satisfaction of target users groups, the approximate upper and lower confidence limits derived for each group of users show that there are no significant difference among the experts' users opinions – almost all intervals in Fig.26 are overlapping. The real difference

can be stated only between opinions of experts groups and the general users. The general users are more satisfied with the ENRICH project results than the experts – content providers, technical staff and scholars.

The results of the quality evaluation performed on the base of a total sample of 205 respondents expressing their opinion on the achievements of the project ENRICH are summarized in the following three tables.

Table II. ***QUALITY CRITERIA*** *– Estimated Scores and Their Percentages of the Maximum Possible Values*

Criteria	Estimate	%
Interoperability	3,60	90,00
Adaptability	3,31	82,75
Usability	3,04	76,00
Security	3,01	75,25
Multilinguality	2,79	69,75

Table III. ***CATEGORIES*** *– Estimated Scores and Their Percentages of the Maximum Possible Values*

Category	Estimate	%
Objects	3,22	80,50
Process	3,19	79,75
Tools	3,12	78,00
Repository	2,49	62,25

Table IV. ***SATISFACTION of Target Users Groups*** *– Estimated Scores and Their Percentages of the Maximum Possible Values*

Target Group	Estimate	%
General users	3,45	86,25
Scholars	3,21	80,12
Technical staff	3,17	79,30
Content providers	3,09	77,30

The evaluation process was composed of the several actions performed from the April to the December 2009 and the evaluation results were in a permanent change as more evaluation data were obtained. The dynamic of those changes was fixed in the Progress Reports made every a half year of project work. The last evaluation results show a real stabilization of estimates concerning the Criteria, Categories or the Users satisfaction; they added minor changes to results obtained in the previous evaluations. This means that we have consistent estimates of quality in all aspects.

The original methodology for evaluation of quality in digital repository were developed enabling to obtain an universal estimator of quality, independent of the number of criteria used for evaluation or individual opinions of evaluators. The structural model proposed enables the multifaceted aspects to be evaluated, to establish relationships between Categories and Criteria and to have a matrix of those relationships across the set of Criteria / Categories. The approximate confidence intervals of proposed estimators of quality, fitted to each items under investigation, enables to make reliable statistical inferences on differences of estimated

values or their comparability to each other with the fixed confidence level, usually with the 0,95 probability.

The developed methodology is applicable in other situations when a quality in a digital environment has to be evaluated. The steps would be the following:

- **Select the set of Criteria and Categories important to that case, fix a number of Criteria and sub criteria needed in that particular case;**
- **Organize the evaluation process by creating questions for evaluation, invite respondents, define the necessary sample sizes;**
- **Calculate the values of double average statistics corresponding to the items under evaluation;**
- **Construct the confidence intervals for investigated quantities, depending on the number of respondents and the number of questions evaluated;**
- **Derive the statistically based inferences on the quality matters estimated.**

References

[1] Manuscriptorium Digital Library http://www.manuscriptorium.eu (www.manuscriptorium.com) [accessed on 5 December 2009]

[2] **ENRICH** - *European Networking Resources and Information concerning Cultural Heritage* project (2007 - 2009) http://enrich.manuscriptorium.com/ [accessed on 5 December 2009]

[3] MINERVA *Technical Guidelines document*, [accessed on 5 December 2009] http://www.minervaeurope.org/publications/technicalguidelines.htm

[4] Brooke, John (1986) SUS - *A quick and dirty usability scale*, [accessed on 5 December 2009] http://www.usabilitynet.org/trump/documents/Suschapt.doc

[5] *UsabilityNe*t: Usability Resources for practitioners and managers http://www.usabilitynet.org/home.htm; http://www.usabilitynet.org/tools/methods.htm [accessed on 5 December 2009].

[6] Kligiene, Nerute, (2009) *E-Accessibility Marking a Quality of Digital Repository*, Proceedings of 2nd International Multi-Conference on Society, Cybernetics and Informatics v. II, July 10-13, 2009, Orlando, Florida, USA, p.p. 167- 172.

[7] Kligiene, Nerute, *Structural Model for Digital Repository Quality Evaluation in Context of Usage*", Proceedings of eChallenges 2009 Conference, 21-23 October, Istanbul.

[8] *Quality principles for Cultural Websites: a Handbook*, 2005 Minerva Project**,** http://www.minervaeurope.org/userneeds/qualityprinciples.htm [accessed on 5 December 2009]

[9] ***Testing of e-Applications Developed in ENRICH****: Usability, Evaluation of Migration Tool, Personalized Translation, Personalization for Contributors and Users* www.musicalia.lt/sus/ www.musicalia.lt/eta/ [accessed on 5 December 2009]

The List of Figures and Tables

The Prototype of Evaluation

The Quality Criteria for Validating ENRICH Work

The List of Tables

Annex 1-(a). The Interactive Questionnaire (SUS) for Usability Evaluation Applied to ENRICH Project Web Site

http://enrich.manuscriptorium.com

	Strongly disagree				**Strongly agree**
I think that the documents I was searching for are well accessible	1	2	3	4	5
I found the descriptions of documents incomplete and difficult to find them	1	2	3	4	5
I found the various functions in this system were well integrated	1	2	3	4	5
I thought there was too much inconsistency in this system	1	2	3	4	5
I thought the website and whole system was easy to use	1	2	3	4	5
I think that I would need the support of a technical person to be able to use this system	1	2	3	4	5
I would imagine that most people would learn to use this system very quickly	1	2	3	4	5
I found the system very cumbersome to use	1	2	3	4	5
I felt very confident using the system	1	2	3	4	5
I found the system unnecessarily complex, I needed to learn how get going with this system	1	2	3	4	5

Annex 1-(b). The Interactive Questionnaire for WP 3 Evaluation and Results

Evaluate the developed *migration tool* on the basis of the sample data sets you have used (corresponding to the Task 3.1.1 - a)	***Yes***	***No***
1. Is the target data format properly documented?	☐	☐
2. Are the interfaces properly documented?	☐	☐
3. Are the underlying transformation Stylesheets available for download and modification?	☐	☐
4. Are a number of recommendations for migration routes available (e.g. Archive-specific (email contact) and Self-Guided)?	☐	☐

	Failed				***Excellent***
Ranking of the **INTEROPERABILITY / Tools**:	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4

Evaluate the developed *migration tool* on the basis of the sample data sets you have used with respect of tool's *adaptability* (corresponding to the Task 3.1.1 - b)	***Yes***	***No***
1. Are the migration tools pre-licensed for adaptation by others?	☐	☐
2. Are schemas in multiple formats provided for suitable validation of resulting conversions?	☐	☐
3. Are the transformations clear enough that (given sufficient technical knowledge) you would be able to adapt them to your own purposes?	☐	☐
4. Are the formats chosen for the migration case studies suitable?	☐	☐

	Failed				***Excellent***
Ranking of the **ADAPTABILITY / Tools**:	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4

Evaluate the developed *migration tool* on the basis of the sample data sets you have used with respect of usability (corresponding to the Task 3.1.1 - c)	***Yes***	***No***
1. Have you encountered no fatal errors causing loss of work in progress?	☐	☐
2. Does the web interface display properly in your browser?	☐	☐
3. Is the interface intuitive/user-friendly enough?	☐	☐
4. Is the accompanying documentation well-designed and helpful?	☐	☐

	Failed				***Excellent***
Ranking of the **USABILITY / Tools**:	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4

Are you satisfied with your evaluation? The rating is from 0 = <failed or feature don't exist> to 4 = <Excellent>:

	Failed				***Excellent***
INTEROPERABILITY / Tools	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4
ADAPTABILITY / Tools	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4
USABILITY / Tools	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4

If the results acceptable, please click on <Submit>; if not, return back to the questions.

You were invited to evaluate results by: *

Select one:

Submit

See statistics...

The average scores obtained from 39 respondents evaluating WP 3

	WP3-a	WP3-b	WP3-c	WP 3
Providers	3,27	3,2	3	
Technical	3,52	3,82	2,74	
Scholars	4	4	3,5	
General users	4	4	2,5	
Average score	3,50	3,60	2,87	**3.32**

	WP3-a	WP3-b	WP3-c
Interoperability/Tools	3,50		
Adaptability/Tools		3,60	
Usability/Tools			2,87

Annex 1-(c). The Interactive Questionnaire for WP 4 Evaluation and Results

Adaptability / Object	***Yes***	***No***
1. Is the extent of descriptive information that can be created for virtual documents adequate for such feature?	☐	☐
2. Is the possibility to add images and textual information (no audio/video support) adequate for virtual documents?	☐	☐
3. Is the extent of descriptive information that can be created for personal collections adequate for such feature?	☐	☐
4. Do the two ways of how to create collection contents cover all your collection related needs when creating a collection?	☐	☐
	Failed	***Excellent***
Resulting evaluation of the Adaptability / Object:	☒ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	

Adaptability / Tool	***Yes***	***No***
1. Is the personal collections interface well adapted to your needs?	☐	☐
2. Is the virtual documents interface well adapted to your needs?	☐	☐
3. Is the possibility to share personal collections and virtual documents (eg. possibility to include links of collections into your documents, website etc..) well adapted to your needs?	☐	☐
4. Is the documentation of the personal collections and virtual documents features well designed and helpful?	☐	☐
	Failed	***Excellent***
Resulting evaluation of the **Adaptability / Tool**:	☒ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	

Security / Process	***Yes***	***No***
1. Is the authentication procedure based on username/password adequate?	☐	☐
2. Do you agree that a system upgrade leading to a more secure authentication (e.g. using HTTPS protocol to encrypt sent/received information) is not necessary?	☐	☐
3. Do you find existing regular daily server-side backup of content of your virtual docs and personal collections to be adequate your needs?	☐	☐
4. Do you find the maximum inaccessibility time of three days needed due recovery from complete server system crashdown to be adequate?	☐	☐

	Failed				***Excellent***
Resulting evaluation of the **Security / Process**:	☉ 0	☐ 1	☐ 2	☐ 3	☐ 4

Are you satisfied with your evaluation? The rating is from 0 = <failled/feature don't exist> to 4 = <Excellent>:

	Failed				***Excellent***
Adaptability / Object	☉ 0	☐ 1	☐ 2	☐ 3	☐ 4
Adaptability / Tools	☉ 0	☐ 1	☐ 2	☐ 3	☐ 4
Security / Process	☉ 0	☐ 1	☐ 2	☐ 3	☐ 4

If the results seems acceptable, please click on <Submit>, if not, return back to the questions.

You were invited to evaluate results by: *

Select one:

The average scores obtained from 50 respondents evaluating WP 4

	WP4a	WP4b	WP4c	WP 4
Providers	3,43	3,43	2,83	
Technical	3,54	3,23	3,15	
Scholars	3,40	3,70	2,30	
General users	3,25	3,25	3,75	
Average score	3,41	3,40	3,01	**3.27**

	WP4-a	WP4-b	WP4-c
Adaptability / Objects	3,41		
Adaptability/Tools		3, 40	
Security/Process			3,01

Annex 1-(d). The Interactive Questionnaires for WP 5 Evaluation and Results

The Questionnaire for the first evaluation:

M-Tool: Usability (corresponding to the Task 5.1.1 - a)	***Yes***	***No***
1. Have you encountered no fatal errors causing loosing work in process?	☐	☐
2. Have you encountered no browser error alerts or display difficulties?	☐	☐
3. *Javascrip*t dependency: it was no dependency on J*avascript*, affecting negatively your work with the M-Tool?	☐	☐
4. Is the interface intuitive / user friendly enough?	☐	☐
	Failed	***Excellent***
Resulting evaluation of the **USABILITY / Tools**:	☒ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	

M-Tool: Adaptability of Tools (corresponding to the Task 5.1.1 - b)	***Yes***	***No***
1. Are the bibliographic description forms well adapted to documents in your collection?	☐	☐
2. Is the M-tool well adapted to properties of digital documents in your collection?	☐	☐
3. Is the documentation of the M-Tool well designed and helpful?	☐	☐
4. Is the single (English) language interface sufficient for your needs?	☐	☐
	Failed	***Excellent***
Resulting evaluation of the **ADAPTABILITY / Tools**:	☒ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	

M-Tool: Process (corresponding to the Task 5.1.1 - c)	***Yes***	***No***
1. Are the responses of the service satisfying (no time demanding delays experienced)?	☐	☐
2. Is the built-in check of descriptive metadata sufficient?	☐	☐
3. Is the built-in check of structural metadata and data availability sufficient?	☐	☐
4. Is the usage of (saving/re-opening and updating) with XML files well resolved?	☐	☐

	Failed				***Excellent***
Resulting evaluation of the **USABILITY / Process**:	☒ 0	☐ 1	☐ 2	☐ 3	☐ 4

Are you satisfied with your evaluation? The rating is from 0 = <failled/feature don't exist> to 4 = <Excellent>:

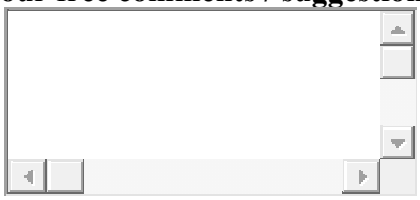
	Failed				***Excellent***
USABILITY / Tools	☒ 0	☐ 1	☐ 2	☐ 3	☐ 4
ADAPTABILITY / Tools	☒ 0	☐ 1	☐ 2	☐ 3	☐ 4
USABILITY / Process	☒ 0	☐ 1	☐ 2	☐ 3	☐ 4

If the results seems acceptable, please click on <Submit>, if not, return back to the questions.

You were invited to evaluate results by: *

Select one:

Your free comments / suggestions:



The average scores obtained from 49 respondents evaluating WP 5 (the first evaluation)

	WP5-a	WP5-b	WP5-c	WP 5
Providers	2,65	3,18	3,59	
Technical	3,17	3,42	3,5	
Scholars	3,33	3,50	3,67	
General users	3,67	3,67	4	
Average score	2,97	3,17	3,60	**3.25**

	WP5-a	WP5-b	WP5-c
Usability/Tools	2,97		
Adaptability/Tools		3, 17	
Usability/Process			3,60

Evaluation and Testing Applied to ENRICH WP5 *NEW (1)*

The Questionnaire for the second evaluation using **existing metadata** exports [New (1)] http://www.musicalia.lt/eta/wp51.php :

The evaluation focuses results achieved during aggregation of documents from each individual content provider (both regular Content Partners and Associated partners).

As there are different ways of cooperation there are different evaluation questionnaires:

- The questionnaire located in this page is dedicated to those who provide **existing metadata** exports for import to Manuscriptorium.
- Another questionnaire is dedicated to those who **create theirs documents metadata** using Manuscriptorium on-line tools.

The questions published at [9] http://www.musicalia.lt/eta/wp51.php focus the interoperability of produced documents, usability of tools and processes and as well their adaptability as experienced during preparation of routine cooperation with each individual content provider.

From each content provider at least one evaluation should be provided. It is suggested to fill the questionnaire separately for each different collection of documents aggregated in Manuscriptorium as the collections have different properties and there are different

connectors (a part of Manuscriptorium import interface) created for each different collection so the results achieved through each the connector may vary.

Each partner is kindly asked to perform the evaluation as soon as receives necessary mappings documentation and XSL transformation. This will be sent to the partners during first days of evaluation. In parallel the documentation and XSL will be published in the project website and within dedicated section of new *Manuscriptorium* beta website.

	WP51-a	WP51-b	WP51-c	WP51-d	WP51-e	WP51-f
Providers	3,64	3,5	3,21	2,5	2,86	2,86
Technical	3,63	2,86	3,26	2,4	3,26	2,8

The average scores obtained from 22 respondents evaluating WP 5 (the second evaluation using **existing metadata** exports [New (1)]**)**

	WP51-a	WP51-b	WP51-c	WP51-d	WP51-e	WP51-f
INETROPERABILITY / Object	3,64					
ADAPTABILITY / Tools		3,23				
USABILITY / Processes			3,23			
USABILITY / Tools				2,55		
ADAPTABILITY / Object					3	
USABILITY / Object						2,82

Evaluation and Testing Applied to ENRICH WP5 *NEW (2)*

The Questionnaire for the second evaluation using *Manuscriptorium* on-line tools [New (2)] http://www.musicalia.lt/eta/wp52.php :

The evaluation focuses results achieved during aggregation of documents from each individual content provider (both regular Content Partners and Associated partners).

As there are different ways of cooperation there are different evaluation questionnaires:

- The questionnaire located in this page is dedicated to those who **create theirs documents metadata** using Manuscriptorium on-line tools.
- Another questionnaire is dedicated to those who provide **existing metadata** exports for import to Manuscriptorium.

The questions published http://www.musicalia.lt/eta/wp52.php focus the interoperability of produced documents, usability of tools and processes and as well their adaptability as experienced during preparation of routine cooperation with each individual content provider.

From each content provider at least one evaluation should be provided. It is suggested to fill the questionnaire separately for each different collection of documents aggregated in Manuscriptorium as the collections have different properties and there are different connectors (a part of Manuscriptorium import interface) created for each different collection so the results achieved through each the connector may vary.

Each partner is kindly asked to perform the evaluation as soon as receives necessary mappings documentation and XSL transformation. This will be sent to the partners during first days of evaluation. In parallel the documentation and XSL will be published in the project website and within dedicated section of new Manuscriptorium beta website.

	WP52-a	WP52-b	WP52-c	WP52-d	WP52-e	WP52-f	Average
Providers	3,88	3,13	3	3	3,38	2,88	3,21
Technical	4	4	4	4	3	4	3,83

The average scores obtained from 9 respondents evaluating WP 5
(the second evaluation using *Manuscriptorium* on-line tools [New (2)])

	WP52-a	WP52-b	WP52-c	WP52-d	WP52-e	WP52-f
INETROPERABILITY / Object	3,89					
ADAPTABILITY / Tools		3,22				
USABILITY / Processes			3,11			
USABILITY / Tools				3,11		
ADAPTABILITY / Object					3,33	
USABILITY / Object						3

	WP5n-a	WP5n-b	WP5n-c	WP5n-d	WP5n-e	WP5n-f	Average
Providers	3,73	3,36	3,14	2,68	3,05	2,86	3,14
Technical	3,67	2,89	3,33	2,78	3,22	2,88	3,13

The average scores obtained from 31 respondents evaluating WP 5
(the second evaluation)

	WP5NEW-a	WP5NEW-b	WP5NEW-c	WP5NEW-d	WP5NEW-e	WP5NEW-f
INETROPERABILITY / Object	3,71					
ADAPTABILITY / Tools		3,23				
USABILITY / Processes			3,19			
USABILITY / Tools				2,71		
ADAPTABILITY / Object					3,1	
USABILITY / Object						2,87

Annex 1-(e). The Interactive Questionnaire for WP 6 Evaluation and Results

Multilingual access via the API integration in the data retrieval interface associated or independent of a multilingual search of digital object (corresponding to the Task 6.1.1 - a)	***Yes***	***No***
1. Is a *time of integration* acceptable?	☐	☐
2. Is the *facility* in customization well designed and helpful?	☐	☐
3. Is the *speed* satisfying?	☐	☐
4. Is the *rendering* acceptable?	☐	☐

	Failed				***Excellent***
Resulting evaluation of the **MULTILINGUALITY / Object:**	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4

Multilingual access via the API integration in the data retrieval interface associated or independent of a multilingual search processing (corresponding to the Task 6.1.1 - b)	***Yes***	***No***
1. Is the *speed* satisfying?	☐	☐
2. Is the *formatting* well resolved?	☐	☐
3. Is the *rendering* acceptable?	☐	☐
4. Is your *interaction* with the system easy to be established?	☐	☐

	Failed				***Excellent***
Resulting evaluation of the **MULTILINGUALITY / Process:**	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4

Multilingual access dedicated **translation interface** where ENRICH expert users can fine-tune dynamically the machine **translation tools** thanks to adapted linguistic tools for terminology extraction and translation post-editing and customization (corresponding to the Task 6.1.2 - a)	***Yes***	***No***
1. Is the interface intuitive/*user friendly* enough?	☐	☐
2. Are you satisfied with the *Quality Assessment*?	☐	☐
3. Is your *interaction* with the translation interface easy to be established?	☐	☐
4. Is the *integration* satisfying?	☐	☐
	Failed	***Excellent***
Resulting evaluation of the **MULTILINGUALITY / Tools:**	☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4	

Multilingual access dedicated **translation interface in repository** (corresponding to the Task 6.1.2 - b)	***Yes***	***No***
1. Are you satisfied with *accessibility* of repository?	☐	☐
2. Is the *interaction* in repository easy to be established?	☐	☐
3. Are you satisfied with the *Quality Assessment in repository*?	☐	☐
4. Is the *Repository Update* satisfying?	☐	☐
	Failed	***Excellent***
Resulting evaluation of the **MULTILINGUALITY / Repository:**	☐ ☐ ☐ ☐ ☐	

	0	1	2	3	4

Are you satisfied with your evaluation? The rating is from 0 = <failled/feature don't exist> to 4 = <Excellent>:

	Failed				***Excellent***
MULTILINGUALITY / Object	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4
MULTILINGUALITY / Process	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4
MULTILINGUALITY / Tools	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4
MULTILINGUALITY / Repository	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4

If the results seems acceptable, please click on <Submit>, if not, return back to the questions.

You were invited to evaluate results by: *

Select one:

Submit

See statistics...

The average scores obtained from 37 respondents evaluating WP 6

	WP6-a	WP6-b	WP6-c	WP6-d	WP 6
Providers	2,95	2,9	2,75	2,45	
Technical	2,83	3	2,5	2,5	
Scholars	3,67	2,67	1,67	1,33	
General users	3,13	3,13	3,13	3	
Average score	3,03	2,95	2,70	2,49	**2.66**

	WP6-a	WP6-b	WP6-c	WP6-d
Multilinguality/Objects	3,03			
Multilinguality/ Process		2,95		
Multilinguality/ Tools			2,70	
Multilinguality/Repository				2,49