



**Cultural and historical digital libraries  
dynamically mined from news archives**

## **Definition of evaluation metrics and tests**

<b>Project Reference No.</b>	<b>FP7-215874</b>
<b>Deliverable No.</b>	<b>D7.2: Definition of evaluation metrics and tests</b>
<b>Workpackage no:</b>	<b>WP7: System testing and user evaluation</b>
<b>Nature:</b>	<b>R (Report)</b>
<b>Dissemination Level:</b>	<b>PU (Public)</b>
<b>Document version:</b>	<b>final</b>
<b>Date:</b>	<b>15/09/2009</b>
<b>Editor(s):</b>	Jan Korsten (SHT), Akrivi Katifori (NKUA/i), Aristorelis Tympas (NKUA/h)
<b>Document description:</b>	This document outlines the way the individual Papyrus modules and the complete Papyrus prototypes will be evaluated.



## History

Version	Date	Reason	Revised by
01	06/08/2008	Table of Contents	J. Korsten
02	03/09/2009	First Draft	J. Korsten
03	09/09/2009	Second Draft after peer review	J. Korsten
04	15/09/2009	Final version	J. Korsten

## Authors List

Organisation	Name
SHT	Jan Korsten, J.W.A.Korsten@tue.nl
NKUA/i	Akrivi Katifori, vivi@di.uoa.gr
NKUA/h	Aristorelis Tympas, tympas@phs.uoa.gr
CINECA	Giorgio Pedrazzi, g.pedrazzi@cinca.it
QMUL	Krishna Chandramouli, krishna.chandramouli@elec.qmul.ac.uk



# Table of Contents

- List of Figures ..... 4
- List of Tables ..... 5
- Executive Summary ..... 6
- 1. Introduction ..... 7
- 2. Functionalities and modules to be tested..... 8
  - 2.1. Papyrus modules to be tested..... 8
  - 2.2. Use cases and functionalities to be tested ..... 8
- 3. Evaluation phases ..... 10
  - 3.1. Testing 1st (provisional) prototype Papyrus ..... 10
  - 3.2. Testing 2nd (final) prototype Papyrus..... 10
- 4. Tests and metrics ..... 12
  - 4.1. User evaluation ..... 12
    - 4.1.1. Ontology Editor ..... 13
    - 4.1.2. Ontology Mapper ..... 13
    - 4.1.3. Ontology Browser and search functionality..... 13
  - 4.2. Content Analysis and Multimedia Retrieval Components ..... 13
    - 4.2.1. ASR evaluation ..... 14
    - 4.2.2. Classification models evaluation ..... 15
    - 4.2.3. Multimedia Retrieval Measures ..... 17
    - 4.2.4. Speaker Diarisation evaluation measures ..... 17
- 5. Conclusions ..... 18
- 6. References ..... 19
- 7. Annex 1 – Basic Questionnaire ..... 20



## List of Figures

Figure 4-1. Diagram illustrating the concepts of recall and precision. Set R is the relevant set from the reference transcription, and set A is the retrieved set from the automatic transcription. Recall is the ratio between the cardinality of the intersection and the relevant set, while precision is the ratio of the cardinality of the intersection and the retrieved set. .... 15

**Figure 4-2 Speaker Diarisation Performance** ..... 17



## List of Tables

Table 2-1. Papyrus use cases to be tested ..... 8

Table 2-2. Papyrus basic functionalities to be tested..... 9

Table 3-1. Institutes which could provide test subjects for the Papyrus evaluation ..... 11

Table 4-1. Summary of metrics to be recorded in the user evaluation ..... 12

Table 4-2. Summary of tests and benchmarks performed on content analysis modules..... 14

Table 4-3 Classification of Documents ..... 16



## Executive Summary

This document outlines the way the individual Papyrus modules and the complete Papyrus prototypes will be evaluated. It lists the functionalities, modules and use cases to be tested and specifies the various evaluation phases. Also the document makes clear which user groups will be involved in the successive evaluation phases. All planned tests and metrics are described. Users will evaluate Papyrus by means of a questionnaire. The developed questionnaire for the user evaluation is attached in the annex.



## 1. Introduction

The aim of Papyrus is to create a cross-disciplinary digital library and show-case it with the domains of News and History. History researchers, either professionals or amateurs will be able to access the primary source content (News multimedia content) through a structured view of the secondary material (history).

In order for Papyrus to be effectively implemented and offer the necessary functionality to meet the aforementioned goals, it should be thoroughly tested. Appropriate tests will ensure that all system components are working properly, both individually and combined in the Papyrus prototype. Furthermore, users should be actively involved in the evaluation of the user interface as well as the quality of support the tool offers to their research.

The user tests will be based on the scenarios presented in D7.1 [1] which are constructed according to the set of functionalities to be tested. The tests serve several objectives:

- Provide feedback to the system developers. This feedback needs to be incorporated in the final Papyrus prototype.
- Showcase the developed framework.
- Show the quality and validity of the Papyrus engine.
- Test the level of satisfaction of the users when employing the prototype.

In order to be able to incorporate the user comments in the final Papyrus prototype the test phase is split in two stages:

During the first stage a selected group of users will test a preliminary, uncompleted version of Papyrus. During a workshop (September 17-18 2009) in Athens two central Papyrus tools, the Mapper and the Browser will be tested. A larger group of users will later test the first Papyrus prototype.

The second phase of the test will involve a larger group of users and all identified user groups.

The user evaluations will result in two reports:

- D7.3a Draft Evaluation report (M24) – based on tests of individual modules and the first Papyrus prototype.
- D7.3b Final Evaluation report (M30) – based on tests of second Papyrus prototype.



## 2. Functionalities and modules to be tested

In order to make sure that Papyrus will accomplish the goals set in the user requirements stage of the project, it is crucial to prepare and implement thorough evaluations of the individual components as well as the whole prototype. These evaluations will include benchmark tests of the components as well as user evaluations.

This section briefly presents the Papyrus modules and the implemented use cases and functionalities that will be tested.

### 2.1. Papyrus modules to be tested

The main focus of the Papyrus evaluation will be in the following Papyrus modules:

- Ontology editor
- Ontology browser and search tool
- Ontology matching and mapping tool
- Content analysis and multimedia retrieval

The testing of these modules will be accomplished either by laboratory benchmarks or through an appropriate user evaluation. The testing will take place in two phases during and after the development of the Papyrus prototype and will continue until the end of the project.

### 2.2. Use cases and functionalities to be tested

Deliverable 7.1 [1] defined 9 use cases that should be tested in the evaluation phase of the project. These are presented in Table 2-1.

**Table 2-1. Papyrus use cases to be tested**

	User Group	Use Case Title	Use Case Description
1	Ontology Administrator	Manage News Ontology	Ontology Administrators may edit the News Ontology and its instances
2	Ontology Administrator	Manage History Ontology	Ontology Administrators may edit the History Ontology and its instances
3	Ontology Administrator	Content Analysis	Ontology Administrators may initiate the automatic analysis of parts of the stored content.
4	Ontology Administrator	Map History and News Ontology entities	Ontology Administrators may create mappings between classes and instances of the history ontology and the news one. This may be done manually or semi-automatically.
5	All	Browse the History and the News Ontology	All users may browse the ontologies and navigate through the mappings between them and also through them to the multimedia content
6	All	Query the History Ontology	All users may perform queries to the History Ontology to get relevant ontology classes and instances as well as multimedia content through them.
7	All	View the search results	All users may navigate within the search results, which may be presented in different ways, according to the needs of the user.

## D7.2: Definition of evaluation metrics and tests



8	All	Save the search results	All users may save the results of the query as well as the query itself for their personal archive
9	End user	Submit new secondary source material	The non-authorized users may prepare and propose new content for the history ontology to be approved by ontology administrators.

**Table 2-2. Papyrus basic functionalities to be tested**

<b>Functionality number D7.1</b>	<b>User category</b>	<b>Basic functionality to be tested</b>
3.2.1.1	Ontology administrator	Add concept
3.2.1.2	Ontology administrator	Add instance
3.2.1.3	Ontology administrator	Edit concept/instance
3.2.1.4	Ontology administrator	Approve/reject submitted material
3.2.1.5	Ontology administrator	Map history and news ontology entities
3.2.1.6	Ontology administrator	Initiate content analysis
3.2.2.1.1	End user	Ontology browsing: Ontology concept view
3.2.2.1.2	End user	Ontology browsing: Topic/subject view
3.2.2.1.3	End user	Ontology browsing: Combined history and news ontology view
3.2.2.1.4	End user	Ontology browsing: Browse ontology within a time period
3.2.2.2.1	End user	Ontology querying: Keyword querying
3.2.2.2.2	End user	Ontology querying: Querying using predefined query types
3.2.2.3	End user	Save retrieved results
3.2.2.4	End user	Submit new secondary source material



### 3. Evaluation phases

#### 3.1. Testing 1st (provisional) prototype Papyrus

The first phase of the testing will take place in the period M18-M24 (September 2009 – February 2010). This testing will result in a draft Evaluation report (D7.3a-M24), which will report on the draft results from the user validation of the individual modules, as well as results from the benchmark tests of individual components and modules. This testing needs to provide information that can be used to further improve Papyrus. It is necessary to identify errors and areas that need improvement. As the system will not be fully complete during this testing phase, not all user groups will be involved in this test.

The modules to be included in the test are:

- Ontology browser and search tool
- Ontology editor
- Ontology matching and mapping tool
- Content analysis and multimedia retrieval

Although the emphasis during this phase lays on the individual modules and components, the aim is also to test the first working version of the entire Papyrus system, provided that a working version of Papyrus will be available in time. Please note that some of the functionalities will be tested more thoroughly in the second phase as they will not be completed in time for the first testing phase.

The first phase of the user testing will proceed in two stages:

The first stage is a workshop that will be organized in Athens (September 17-18) with the following objectives:

- (PhD) students will define the mappings between news and history ontology by using the Papyrus Mapper and the Papyrus Browser.
- The selected group of users will provide feedback on the tools used.
- A discussion between users and system builders will take place on multilingual issues.

The second stage of this testing phase will be to evaluate the first Papyrus prototype as a whole and not just the individual tools. This will involve ontology administrators and end users, which at this stage will be at the level of historical researchers and students (advanced and intermediate users). Amateurs will be involved in the second testing phase. The scenarios to be used in the end user evaluation will be based on the ones defined in D7.1 [1]. The final scenarios depend also on the state of the art of the first Papyrus prototype.

This testing phase will result in the preliminary evaluation report D7.3a (M24).

In order to record the user responses a questionnaire has been developed. This questionnaire contains a set of basic questions and data. This basic questionnaire forms the basis for the questionnaires that will be used during the different user tests. Other instruments used during the test are benchmark tests and the recording of user actions.

More details on the tests are available in section 4.

#### 3.2. Testing 2nd (final) prototype Papyrus

The second and final testing and evaluation phase is scheduled for the period M24-M30 (February 2010-August 2010). This phase contains user tests of the entire Papyrus prototype, the processing and analyzing of the test results; and the composition of the final evaluation report D7.3b.

## D7.2: Definition of evaluation metrics and tests



This testing phase will evaluate the whole Papyrus prototype as all its modules are expected to be fully working and incorporated by then. It will be a user evaluation and will involve ontology administrators as well as all types of end users: amateur, intermediate and advanced.

Testing will take place in different settings. They will be determined in M24, taking into account the experiences of the first testing phase. Possible settings are:

- Bringing different user groups together in a number of central locations where they can access and use Papyrus. In this setting it is possible to provide sufficient instructions. User reactions can be recorded and the users can be interviewed afterwards.
- Making Papyrus accessible via the internet (using usernames and passwords). This makes it possible to get more user responses. They can evaluate the system by using an online evaluation form, for instance by using SurveyMonkey.com.

The scenarios to be used in the end user evaluation will be based on the ones defined in D7.1 [1]. These will be refined following the experience of the first phase of the evaluation.

Part of the testing is also to record a number of benchmarks.

In order to record the user responses a questionnaire has been developed. This questionnaire contains a set of basic questions and data. This basic questionnaire forms the basis for the questionnaires that will be used during the different user tests. Other instruments used during the test are benchmark test and the recording of user actions.

More details on the tests are available in section 4.

Table 3-1 gives an overview of the users foreseen to be involved in the two testing phases. For the first phase of the evaluation, it is expected that about 10 users, all researchers and students, will take part in the workshop and another 10-15, again mostly researchers and students as well as 3-5 journalists will participate in the laboratory evaluation.

For the second phase, the final Papyrus evaluation, 10-15 users from each level of expertise (amateur, intermediate and advanced) will be involved.

**Table 3-1. Institutes which could provide test subjects for the Papyrus evaluation**

Users	Institute
PhD students	Foundation for the History of Technology (located in Prague and Plovdiv)
	Eindhoven University of Technology
	National and Capodistrian University of Athens
	Tensions of Europe network
Senior researchers	Eindhoven University of Technology
	National and Capodistrian University of Athens
	Tensions of Europe network
Journalists	Deutsche Welle
	AFP
University Students	Eindhoven University of Technology,
	NKUA
	European University Institute Florence
Amateurs	Option: historical section Dutch Organization of engineers
	Option: volunteers several (Dutch) amateur historical associations

## 4. Tests and metrics

This section presents the benchmark tests as well as the usability evaluations that will be performed during the evaluation of Papyrus.

Section 4.1 presents the method for user evaluation that will be applied for the Papyrus components which the users will be directly involved with, namely the Editor, the Mapper and the Browser.

### 4.1. User evaluation

**Table 4-1. Summary of metrics to be recorded in the user evaluation**

Metric	Tool for recording the metric
Correct, wrong, incomplete and no answer percentages	Recorded during the experiment
Time for task completion	Recorded during the experiment
User satisfaction	Evaluation questionnaire and interview
User comments and problems	Recorded during the experiment and discussed in the interview

For the user evaluation we decided to proceed with a scenario-based usability test accompanied with a usability questionnaire, to be performed in a laboratory with the cooperation of the users of the involved user groups. To prepare this experiment we took into account existing proposed methodologies for evaluating similar web-based tools [2] following NKUA's prior experience with scenario-based usability testing [3].

Furthermore, the usability questionnaire compiled will be distributed to individuals involved in using the Papyrus platform to either perform historical research or edit the ontology and the mappings in real use of the tools and not in an organized laboratory session.

The evaluation will proceed in the following stages:

**Training.** Before the beginning of the evaluation, about thirty minutes for each user will be dedicated on explaining the use and concepts related to the tools to be evaluated.

**Main experiment.** After the training period, users are asked to perform a set of tasks designed to test the individual tools. These tasks will be designed based on the use cases and functionality/functionality to be tested (Table 2-1 and Table 2-2) as well as the general user scenarios defined in [1].

For each task the user has to perform a certain function (editing the ontology, searching, etc) given by the experimenters. The user has to conclude the tasks in the sequence in which they are given. There is a time limit for each task and the participants are not allowed to backtrack to an earlier task.

After completing the tasks the users are asked to fill in a questionnaire that is derived from the basic questionnaire (Section 7 - Annex) in order to record their reaction to the tool. This questionnaire contains closed-ended questions (i.e. questions where the respondent should choose among a fixed set of responses) as well as open-ended ones (questions for which the respondent is asked to provide free text, allowing her/him to provide comments and feedback on the tool). The questionnaire contains general questions that are relevant to all the evaluated tools and specific ones for each tool.

During the experiment the **time** needed for a user to complete a task is recorded. His/her failure to complete the task is recorded as well, along with any comments or reactions and difficulties that the user may have with certain tasks. Subjects will be asked to think aloud in order to record any comments on the tools.

It should be noted that when a participant completes a task, the answer or successful task completion will be noted down, and in the analysis stage the following percentages will be computed:

- Correct answers: Percentage of correct answers given.
- Wrong answers: Percentage of wrong answers.
- Incomplete answers: Answers that are correct but did not contain all the requested information or completed part of the task.
- No answer: This percentage refers to the queries that participants choose not to answer because they didn't know how to proceed further or did not want to pursue the task further.

### ***4.1.1. Ontology Editor***

Ontology administrators will be involved in this test. They will be asked to perform certain tasks that will involve making specific changes in the ontology.

### ***4.1.2. Ontology Mapper***

As before, ontology administrators will be involved in this test. They will be asked to perform certain tasks that will involve mapping concepts from the 2 ontologies.

### ***4.1.3. Ontology Browser and search functionality***

End users will be involved in this test. User tasks for this test will include browsing to locate specific entities or news items. Precision and recall of the retrieved results should be measured as well.

In the case of the search functionality, comparative tests will be planned to evaluate searching through the history ontology against keyword search directly on the news items through Lucene search engine.

Table 4-1 provides a summary of metrics to be recorded for every Papyrus module participating in the user evaluation. Through these metrics possible problems as well as the usability of the Papyrus prototype will be determined.

## **4.2. Content Analysis and Multimedia Retrieval Components**

For the evaluation of content analysis and multimedia retrieval components thorough testing has been and will applied. The results will be presented, along with the results in the respective deliverables of WP4 – Targeted Multimedia Content Analysis.

This section presents briefly the tests and measures to be used.

Table 4-2. Summary of tests and benchmarks performed on content analysis modules

Component	Test/benchmarks	Main Metrics
Speaker Diarisation	Nist benchmark/manual segmentation of a subset of Papyrus audio	Diarisation Error Rate
ASR	Text Transcriptions of 5 Deutsche Welle video	Word Error Rate, Precision, Recall, F-score
Text classification	Papyrus and external documents already classified	Accuracy, Precision, Recall, F-score
Metadata extraction (under evaluation)	Manually Annotated Papyrus Documents	To be decided
Multimedia Retrieval (and Relevance feedback)	The list of multimedia documents annotated as relevant and irrelevant for a given query	Precision-Recall curve

#### 4.2.1. ASR evaluation

The standard evaluation metric for speech recognition systems is the **word error rate** (WER). The word error rate is based on how much the word string returned by the recognizer (often called the **hypothesized** word string) differs from a **reference** transcription. Given such a correct transcription, the first step in computing word error is to compute the **minimum edit distance** in words between the hypothesized and correct strings. The result of this computation will be the minimum number of word **substitutions**, word **insertions**, and word **deletions** necessary to map between the correct and hypothesized strings. The word error rate (WER) is the average number of word recognition errors per reference word (note that because the equation includes insertions, the error rate can be greater than 100%):

Word Error Rate =  $100 \times \frac{\text{Insertions} + \text{Substitutions} + \text{Deletions}}{\text{Total Words in Correct Transcript}}$

The **Word Recognition Rate** is defined as  $WRR = 1 - WER$ .

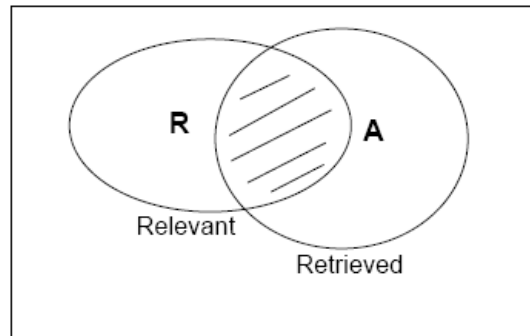
More high level metrics such as sentence error rate and concept error rate can be applied, depending on the applications. The **Sentence Error Rate** tells how many sentences had at least one error:

Sentence Error Rate =  $100 \times \frac{\# \text{ of sentences with at least one word error}}{\text{total } \# \text{ of sentences}}$

The **concept error rate** is useful when a Natural Language Understanding component is involved to produce a semantic representation of the speech. It measures the percentage of semantic concepts that the system returns correctly (e.g. flight destination and departure time).

In less definite tasks, such as open-domain speech understanding, the **weighted keyword error rate** (WKER) has been proposed, to avoid keywords and functional words being treated in the same manner. The WKER is based on the tf-idf criterion used in information retrieval.

An alternative evaluation framework is to consider the speech recognition evaluation as an information retrieval task, in which each word occurrence is treated as a unit of information, and in which the goal is for the relevant information present in the reference transcription to be retrieved in the automatic transcription. Measures most commonly used to evaluate information retrieval are the recall and precision.



**Figure 4-1. Diagram illustrating the concepts of recall and precision. Set R is the relevant set from the reference transcription, and set A is the retrieved set from the automatic transcription. Recall is the ratio between the cardinality of the intersection and the relevant set, while precision is the ratio of the cardinality of the intersection and the retrieved set.**

**Recall** is the fraction of the relevant information units (set  $R_i$ ) which has been retrieved:

$$\rho_{\mu} = \frac{\sum |R_i \cap A_i|}{\sum |R_i|}$$

**Precision** is the fraction of the retrieved information units (set  $A_i$ ) which is relevant:

$$\pi_{\mu} = \frac{\sum |R_i \cap A_i|}{\sum |A_i|}$$

Recall corresponds to the number of correctly recognized words over the total words in the reference transcription, precision to the number of correctly recognized words over the total words in the automatic transcription.

While calculating both recall and precision measures offers the most flexible basis for performance analysis, it may sometimes be desirable to evaluate or optimise a system in terms of a single measure. The recall and precision measures can be combined in a single value in a number of ways. One common such measure is the **F-measure**, which is the harmonic mean of recall and precision:

$$F_{\mu} = \frac{(2\pi_{\mu}\rho_{\mu})}{(\pi_{\mu} + \rho_{\mu})}$$

This measure corresponds the closest to the word recognition rate, as it measures the performance over an entire word sequence, with each word occurrence being weighed equally.

#### 4.2.2. Classification models evaluation

Three of the most widely used models to evaluate the accuracy of each classification model are:

- k-fold cross-validation,
- Leave-one-out cross-validation,
- Repeated random sub-sampling validation.

There are various performance measures to determine effectiveness; however, precision, recall, and accuracy are the most often used. To determine these, one must first begin by understanding if the

## D7.2: Definition of evaluation metrics and tests



classification of a document was a true positive (TP), false positive (FP), true negative (TN), or false negative (FN).

**Table 4-3 Classification of Documents**

TP	Documents being classified correctly as relating to a category.
FP	Documents that are erroneously related to the category.
FN	Documents that are not marked as related to a category but should be.
TN	Documents that should not be marked as being in a particular category and are not.

Accuracy is the most commonly used as evaluation measure for categorization techniques and is the percentage of correctly classified documents over the total number of documents.

$$A_i = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}$$

When there are few instances of the interesting category, the overrepresentation of the negative class can cause problems in evaluating classifiers performances using accuracy. Since accuracy is not a good metric for skewed datasets, the classification performance of algorithms in this case is measured by precision and recall.

Precision  $\pi_i$  is determined as the conditional probability that a random document  $d$  is classified under  $c_i$  or what would be deemed the correct category. It represents the classifiers ability to place a document under the correct category as opposed to all documents placed in that category, both correct and incorrect:

$$\pi_i = \frac{TP_i}{TP_i + FP_i}$$

Recall  $\rho$  is defined as the probability that, if a random document  $d_x$  should be classified under category ( $c_i$ ), this decision is taken.

$$\rho_i = \frac{TP_i}{TP_i + FN_i}$$

For obtaining overall estimates of precision and recall, two different methods may be adopted:

- *microaveraging*: precision and recall are obtained by summing over all individual decisions,
- *macroaveraging*: precision and recall are first evaluated "locally" for each category and then "globally by averaging over the results of the different categories.

Furthermore, precision and recall are often combined in order to get a better picture of the performance of the classifier. This is done by combining them:

$$F = \frac{2\pi\rho}{\pi + \rho}$$

where  $\pi$  and  $\rho$  denote precision and recall respectively.



### 4.2.3. Multimedia Retrieval Measures

In conventional multimedia retrieval systems, the retrieved media objects are labelled as either positive or negative providing feedback to improve the retrieval performance. In order to evaluate the automatic decision from a machine learning technique a structure known as a confusion matrix or contingency table will be used. The confusion matrix has four categories as mentioned in Section 4.4.2.

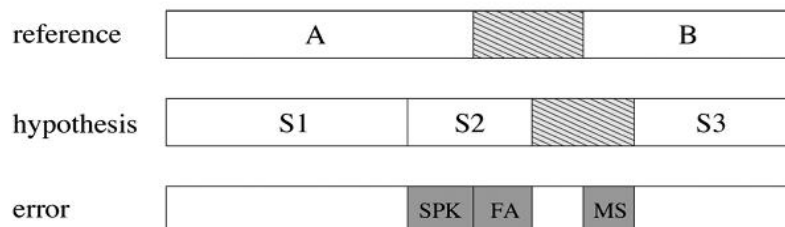
Precision-Recall (PR) curves, are commonly used in Information Retrieval (Manning et al 1999), have been cited as an alternative to ROC curves for tasks with a large skew in the class distribution (Bockhorst et al 2005). An important different between ROC space and PR space is the visual representation of the curves. Study of PR curves can expose differences between algorithms that are not apparent in ROC space. ROC and PR curves are typically generated to evaluate the performance of a machine-learning algorithm on a given dataset.

However, in both cases (ROC space or PR space) each point represents a specific machine learning algorithm output with a threshold for calling an example positive. Thereby, the convex hull can be seen as constructing a new ranked list, as one picks the best points. Therefore, it would be methodologically incorrect to construct a convex hull or achievable PR curve by looking at performance on the test data and then constructing a convex hull. This measure will be used to evaluate the multimedia retrieval algorithms developed in Papyrus system.

### 4.2.4. Speaker Diarisation evaluation measures

The output of a Speaker Diarisation system consists of metadata describing speech segments with starting time, ending time, and speaker cluster name. In order to test the system performance, its output must be compared to a manually annotated ground truth. One of the most used metric to evaluate the performance of these systems is called Diarisation Error Rate (DER).

Since the hypothesis speaker labels may differ from those reported in the reference file, a one-to-one mapping of the reference speaker IDs to the hypothesis speaker IDs must be performed before the evaluation. The algorithm employed maximizes the total overlap of the reference and corresponding hypothesis speakers. Speaker diarisation performance is then expressed in terms of the missed speech (MS) (speech parts not assigned to any speaker), false alarm (FA) (non-speech parts assigned to a speaker), and speaker-error (SPK) (mislabelled speaker) rates, as shown in figure 4.2 The overall diarisation error rate (DER) is the sum of these three components (scored by percentage of total time).



$$DER = \text{Speaker Error (SPK)} + \text{False Alarm Speech (FA)} + \text{Missed Speech (MS)}$$

Figure 4-2 Speaker Diarisation Performance



## 5. Conclusions

In this document we specified the main evaluation metrics and tests. Two kinds of tests have been determined:

- Technical benchmark tests to establish information about the performance and effectiveness of Papyrus.
- User tests that are necessary to retrieve information about user satisfaction and the quality of the user interface. It is expected that tests and metrics will be further refined as following the tests of the individual modules and the further development of the Papyrus prototype. This will be determined by the advances made in the other work packages.



## 6. References

- [1] Papyrus project deliverable D7.1 Domain Specific Validation Scenarios
- [2] A Grani, V Glavini, S Stankov, Usability Evaluation Methodology for Web-based Educational Systems'- 8th ERCIM Workshop: User Interfaces for All, 2004
- [3] A. Katifori, E. Torou, C. Vassilakis, G. Lepouras, C. Halatsis, *Selected Results of a Comparative Study of Four Ontology Visualization Methods for Information Retrieval tasks*, Proceedings of IEEE RCIS 2008



## 7. Annex 1 – Basic Questionnaire



### Basic Questionnaire for user evaluation

*Papyrus is a project that intends to showcase the use of a cross-discipline digital library for the recovery of history from news digital content. It will attempt to understand user queries in the context of the History discipline, look for content in the News domain and return the results in a way useful to the user.*



This questionnaire is purposed to record user satisfaction with the Papyrus platform in order to identify its usability. User reactions and ideas are essential in order to correct errors and make Papyrus efficient and usable.

**Respondent' S Details**

---

\*Name:

\*Organization:

Profession:

Country:

\* Optional fields



## 1. Personal Information

---

To help us evaluate your answers, please indicate:

### 1.1. Your age is between

- 11-17
- 18-25
- 26-35
- 35-50
- 50-

### 1.3. Your gender

- Male
- Female

### 1.3. Your education level

- Highschool
- BA – College Degree
- MA – MsC Degree
- PhD Student
- PhD Holder
- Other (Please indicate .....)

### 1.4. Your interest in History is

- Professional
- Amateur

### 1.5. What languages have you mastered?

- English

**D7.2: Definition of evaluation metrics and tests**



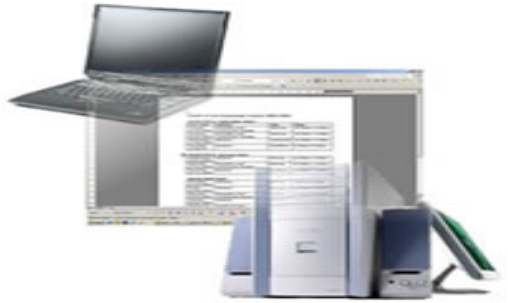
- German
- French
- Italian
- Spanish
- Other (Please indicate)

**1.5. What languages do you use in your research?**

	<b>ALWAYS</b>	<b>OFTEN</b>	<b>NEVER</b>
<b>English</b>			
<b>German</b>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<b>French</b>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<b>Italian</b>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<b>Spanish</b>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<b>Other (please specify)</b>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

.....

## 2. ICT skills



**2.1 How many years have you been using a computer ?**

- 1 year or less
- 2 to 4 years
- 5 years or more

**2.2 Do you use a computer for your work?**

- Yes
- No

**2.3 Do you use a computer for your studies?**

- Yes
- No

**2.4 Do you have a computer at home?**

- Yes
- No

**2.5 How much time do you spend with a computer for your work?**

- Under 1 hour per day
- 1-3 hours a day
- 3-6 hours a day
- More than 6 hours per day
- Not applicable

**2.6 How much time do you spend with a computer for leisure?**

- Under 1 hour per day

## D7.2: Definition of evaluation metrics and tests



- 1-3 hours a day
- 3-6 hours a day
- More than 6 hours per day
- Not applicable

### 2.7 How much time do you spend with a computer for education?

- Under 1 hour per day
- 1-3 hours a day
- 3-6 hours a day
- More than 6 hours per day
- Not applicable

### 2.8 How comfortable do you feel using Computers, in general?

- Very comfortable
- Comfortable – I can do most things I want to do
- Neither comfortable nor uncomfortable
- Uncomfortable



## D7.2: Definition of evaluation metrics and tests



- Very uncomfortable
- Not applicable

### 2.9 How often do you use the internet?

- I do not use the internet
- Under 1 hour per day
- 1-3 hours a day
- 3-6 hours a day
- More than 6 hours per day

### 2.10 How comfortable do you feel using the Internet?

- Very comfortable – I can do everything that I want to do
- Comfortable – I can do most things I want to do
- Neither comfortable nor uncomfortable
- Uncomfortable – I can not do many things I would like to do
- Very uncomfortable - I can not do most things I would like to do
- Not applicable



### 3. Researcher Profile

#### 3.1 What type of user are you?

- Professional user: History Researcher
  - University Student
  - Amateur
  - Professional user: Journalist
  - Professional - Other
- Other: .....

Comments: .....

#### 3. Why do you search historical information?

	ALWAYS	OFTEN	NEVER
Write a dissertation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Write an academic paper	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Write a journalistic article	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Prepare a TV programme	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Prepare a radio programme	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Professional work	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<b>(other than history or journalistic)</b>			
Entertainment/curiosity	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Write a student term paper	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Other (please specify)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
.....	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
.....	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>



3.3 Please indicate some of the topics you have investigated in the past?

.....

3.6. What type of sources do you employ<sup>1</sup>?

	ALWAYS	OFTEN	NEVER
Primary	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Secondary	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

3.7 What kind of sources do you use for your research (multiple options possible):

	ALWAYS	OFTEN	NEVER
Traditional archives	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Digital archives	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Video fragments	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Newspapers and magazines	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Audio fragments	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Books	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Images (photos/paintings)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Oral sources	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

---

<sup>1</sup> \* A primary source is an original source of information that has been created by an authoritative individual with direct knowledge of the fact he/ she describes or impresses. On the contrary, secondary source discusses information originally presented elsewhere (primary sources or other secondary sources). For example, an article in the history of stem cells in the journal *History of Biology* is a secondary source for a historian who researches the history of stem cells by using as primary material references on stem cells found in news agencies items.



## 4. Use of the Web

### 4.1. Do you use the Web for acquiring information

regarding your research /work /studies, for example through a search engine, like Google?

YES

NO because:

- I prefer printed material /sources
- I do not trust the internet
- I do not know any interesting online services
- I find using the internet search engines complicated /troublesome

Comments: .....

### 4.2. If yes, please indicate which search engines do you use:

.....  
.....  
.....  
.....

#### 4.2.1. If yes: indicate how you got to know the search engines you use?

- It was part of the curriculum of your highschool/university
- Colleagues told me about it
- Accidentally
- Other

### 4.3. Have you ever visited a Digital Library or Archive through the Web?

**D7.2: Definition of evaluation metrics and tests**



- YES
- NO

Comments: .....

**4.4. If yes, please indicate which one(s):**

.....  
.....  
.....  
.....

**4.5 What did you do with the retrieved and for your research useful material?**

- I stored it in my own computer
- I stored it on a storage facility provided by the search engine/digital archive
- I printed it and kept no digital copy
- I photographed it
- I just read it and took notes (no storage of actual documents).
- Other .....



## 5. Usability Questionnaire

### 5.1. The tool was:

#### a. interesting

	1	2	3	4	5	6	7			0
disagree								agree	NA	

#### b. usable

	1	2	3	4	5	6	7			0
disagree								agree	NA	

#### c. pleasant

	1	2	3	4	5	6	7			0
disagree								agree	NA	

#### d. effective

	1	2	3	4	5	6	7			0
disagree								agree	NA	

#### e. frustrating

	1	2	3	4	5	6	7			0
disagree								agree	NA	

#### f. confusing

	1	2	3	4	5	6	7			0
disagree								agree	NA	

#### g. tiresome

	1	2	3	4	5	6	7			0
disagree								agree	NA	

### 5.2. I feel in control when I'm using this tool.

	1	2	3	4	5	6	7			0
disagree								agree	NA	

**5.3. This tool uses terms understandable and familiar to me.**

	1	2	3	4	5	6	7			0
disagree								agree	NA	

**5.3. This tool needs more introductory explanation.**

	1	2	3	4	5	6	7			0
disagree								agree	NA	

**5.4. I find this tool useful .**

	1	2	3	4	5	6	7			0
disagree								agree	NA	

**5.5. Everything on this tool is easy to understand.**

	1	2	3	4	5	6	7			0
disagree								agree	NA	

**5.6. This tool is too slow.**

	1	2	3	4	5	6	7			0
disagree								agree	NA	

**5.7. I get what I expect when I click on objects on this tool.**

	1	2	3	4	5	6	7			0
disagree								agree	NA	

**5.8. It is difficult to move around in the tool .**

	1	2	3	4	5	6	7			0
disagree								agree	NA	

**5.9. I feel efficient when I'm using the tool.**

	1	2	3	4	5	6	7			0
disagree								agree	NA	

**5.10. Compared to what I expected, I finished my tasks really quickly**

	1	2	3	4	5	6	7			0
--	---	---	---	---	---	---	---	--	--	---

**D7.2: Definition of evaluation metrics and tests**



disagree									agree	NA	
----------	--	--	--	--	--	--	--	--	-------	----	--

**5.11. I would characterize the tool as innovative.**

	1	2	3	4	5	6	7			0	
disagree									agree	NA	

**5.12. It was easy to learn how to use the tool.**

	1	2	3	4	5	6	7			0	
disagree									agree	NA	

**5.13. Overall, I'm satisfied with the tool.**

	1	2	3	4	5	6	7			0	
disagree									agree	NA	

**5.14. I would use this tool for my research.**

	1	2	3	4	5	6	7			0	
disagree									agree	NA	

**5.15. Please record the positive sides of the tool.**

.....  
 .....

**5.16. Please record the negative sides of the tool.**

.....  
 .....

**5.17. Please comment on how the tool could be improved.**

.....  
 .....