



Spatially-Aware Information Retrieval on the Internet

SPIRIT is funded by EU IST Programme
Contract Number: IST-2001-35047



Evaluation Methodology

Deliverable number:	D19 7201
Deliverable type:	R
Contributing WP:	WP 7
Contractual date of delivery:	1 July 2004
Actual date of delivery:	30 June 2004
Authors:	Guillaume Aumaitre, Sandrine Balley, Stephen Levin
Keywords:	Evaluation, Methodology, Usability, Performance

Abstract: This document describes the general methodology used by the SPIRIT consortium to evaluate SPIRIT interim and final prototypes.

Contents

- 1.INTRODUCTION..... 4**

- 2.EXISTING EVALUATION METHODS..... 5**

 - 2.1.Evaluation of a search engine’s performance..... 5**
 - Evaluation campaigns..... 5

 - 2.2.Evaluation of software usability..... 7**
 - Usability indicators..... 7
 - Testing methods..... 9

- 3.EVALUATION TESTS FOR SPiRiT..... 11**

 - 3.1.Evaluation context..... 11**

 - 3.2.Stability test..... 11**
 - Methodology..... 11

 - 3.3.Global test for system usability and performance..... 11**
 - Methodology..... 11
 - Analysis of results..... 13

 - 3.4.Comparative test..... 14**
 - Methodology..... 14
 - Analysis of results..... 14

- 4.CONCLUSION..... 14**

- 5.REFERENCES..... 15**

Executive Summary

This report describes the methodology that will be used to evaluate the interim SPiRiT prototype and the final SPiRiT prototype. It is composed of several methods. Depending on the prototype's functionalities, these methods may be partly applied. Moreover, these methods will possibly be improved if some insufficiencies are pointed out during the evaluation of the interim prototype.

After an introductory section, the first part of the report describes existing evaluation methods for search engines and for software systems in general. The second part of the report describes the evaluation methodology adopted for use in SPiRiT that is composed of several tests .

D19 7201

Evaluation Methodology

2. Introduction

Quality management provides guidelines for evaluation. More specifically, the task of software evaluation has been described in the application of the general ISO 9001 standard for quality management [ISO 9001 1994] for the specific case of computer software [ISO 9000-3 1997]. Evaluation consists in comparing what has actually been done with what was intended to be achieved. In software engineering, the evaluation process should meet each part of the system conception lifecycle, as shown in the “V-cycle” of software engineering (Figure 1).

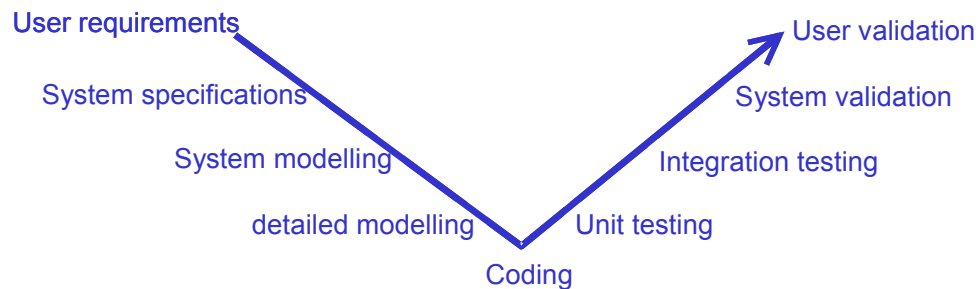


Figure 1: Software engineering cycle (Standards for quality management ISO 9001)

According to this figure, SPIRIT’s evaluation methodology may be organised into four parts.

- Unit testing

Each part of the system (interface, search engine core, ontology, spatial ranking...) was designed separately and has to be tested after being implemented and before integration.

- Integration testing

Integration testing checks that all parts of the system can work together.

- System validation

The system validation step intends to assess that the system matches its design specifications. Some global performance testing can be carried out during this step.

- User validation

The purpose of user validation is to assess the degree to which the identified user requirements have been met, and to provide a feedback in a form that can be used by designers and developers for further developments.

Unit testing and integration testing are carried out in WP1-D16-1101 [Finch et al., 2004]. The system validation step is partly dealt with in WP1-D16 and in this deliverable (see section 3.1): WP1-D16 provides measurements of the effectiveness of particular options combinations

inside the system. In this deliverable, a method is proposed to evaluate the performance of the information retrieval functionality with fixed settings.

The main focus of this deliverable is "user validation" as described in section 3.2.

3. Existing evaluation methods

The ISO 9001 standard provides very general guidelines for evaluation. More specific methods are proposed in the domain of software usability testing that we may use to plan user validation in SPiRiT. These methods are exposed in section 3.2. Some other specific methods take place in the domain of evaluating search engines performance and we will use them to plan system validation in SPiRiT. These methods are discussed in the following section.

3.1. Evaluation of a search engine's performance

Regarding system validation, we limit ourselves to evaluating the search engine performance and more precisely the results of its functionality. The methods and measurement usually employed are part of very well-defined testing frameworks.

Evaluation campaigns

In order to evaluate and to compare Information Retrieval Systems (IRS), some international campaigns have been organized for several years. The Text REtrieval Conferences¹ (TRECs) and Amaryllis² are two such campaigns that allow to test systems on large scale. The TRECs, sponsored by the Defense Advanced Research Projects Agency (DARPA) and the National Institute of Standards and Technology (NIST), have been held every year since 1992. The Amaryllis campaign was organized by INIST (Institut de l'Information Scientifique et Technique) and sponsored by AUF (Agence Universitaire de la Francophonie) and by the French Ministry of Research. The Amaryllis cycle draws heavily on TREC methodology.

Both evaluation campaigns are based on a comparison of performance between each IRS. [Lespinnasse et al., 1998]

Two indicators are usually employed during these campaigns.

The Recall, which is the ratio between the number of relevant docs retrieved by the IRS and the total number of relevant documents among all the documents indexed, may be seen as the probability that a relevant document is retrieved [Loupy et al., 2000].

The Precision, which is the ratio between the number of relevant documents retrieved by the IRS and the total number of documents indexed, may be seen as the probability that a retrieved document is relevant [Loupy et al., 2000].

To measure those two indicators, an evaluation campaign uses a set of 3 collections that is called a test reference collection set [Van Zwol, 2002]:

- corpus "data": a set of documents
- corpus "topics": a set of themes
- corpus "relevance judgments" (also called "Qrels"): the lists of documents among the corpus data that are relevant for each theme in the corpus topics.

A campaign runs in two steps: a training session and a test session.

During the training session, engineers from the various evaluated IRS are given a comprehensive set of test collections they can use to calibrate their systems [Lespinnasse et al., 1998].

¹Trec.nist.gov

²www.upmf-grenoble.fr/adest/seminaires/inist

Then a new test reference collection is defined (e.g. the ad hoc test of TREC uses a new corpus “topics” with the corpus “data” that was already used for the training session but with a new corpus “relevance judgements”). During the test session, the engineers are given the new corpus data and topics on which they have to run their system. The retrieved documents are then compared with new corresponding corpus relevance judgments. This comparison gives measures of recall and precision for each search engine.

A key step in these campaigns is the production of the Qrel, as described below.

In TREC, the document collections (corpus data) used are very large (more than 1Gb) so they can simulate the web. Human assessors can't judge every document in the collection. Therefore, a pooling method is used: the top 100 documents from each of the IRS participating in TREC are pooled and redundant documents are removed. Then, the final set of documents is manually examined for relevance. In the case of TREC and Amaryllis, a binary relevance is used (each document can only be judged as being relevant or not relevant). The pooling method implies that the final list of relevant documents is not exhaustive and that the relevant documents depend on the number of IRS and their performance. This implies that a new evaluation using old TREC experiments is only a comparison with systems that participated in this TREC session. Indeed, if a new method implemented by a new IRS is employed, this new system can retrieve relevant documents that were not in the relevance judgements, so it is impossible to measure the contribution of the new method.

In Amaryllis, the pooling method is not used. Compared with TREC, the Amaryllis corpus are smaller, that is why the lists of relevant documents used for the evaluation are manually built by archivists, and then revised according to the top answers obtained by the participants.

However it is difficult to utilise both of the above methods in small scale projects. Firstly, most of the time, the only available IRS is the evaluated one, which prevents the use of the pooling approach. A second reason is that manual assessment of the corpus is a time consuming process which may be beyond the resources of small scale projects.

Other methods that reduce the effort in creating test collections have also been explored, although they are not currently used in TREC. The most interesting method is the Interactive Searching and Judging method (ISJ) [Cormack et al., 1998]. For each topic a searcher is instructed to search as many variations and refinements of the topic he/she can think of, noting all relevant documents retrieved. Spending on average two hours per topic, the ISJ method is almost as reliable as the TREC pooling method.

3.2. Evaluation of software usability

Several web-resources are available on the topic of usability testing, notably, the following two sites cover the topic of ergonomics, (<http://www.ergolab.net> and <http://www.usabilitynet.org>). We also used some material from the ISO 9241 standard on usability [ISO 9241-11].

These resources provide guidelines, rules and lists of agreed standards for designing “usable” interactive systems, as well as for designing usability tests for evaluating those systems.

Usability indicators

The ISO 9241 standard defines usability as “the effectiveness, efficiency, and satisfaction with which specified users achieve specified goals in particular environments.”

These 3 indicators are detailed below.

Effectiveness means how well the user performs with the system. ISO 9241 defines it as “the accuracy and completeness with which specified users can achieve specified goals in particular environments”. Practically, effectiveness analysis focuses on evaluating whether the

users can use the system's main functionality at all, that is, whether they can perform at all the main user task(s). In SPiRiT, this means: can users find relevant geo-documents.

Efficiency means how efficiently users can work with the system. It is defined by ISO 9241 as "the resources expended in relation to the accuracy and completeness of goals achieved". Practically, efficiency analysis focuses on evaluating how well users can perform their main task(s). Efficiency thus refers to general properties of the main user task like the time required to retrieve information, as well as to a decomposition of this main task into subtasks possibly corresponding to the use of specific functions introduced by the system designer to contribute to the main functionality. These subtasks are, for instance: to learn to work with the system, to interact with a map, to understand the ranking of retrieved results.

Satisfaction means the general user's feeling about the system. It is defined by ISO 9241 as "the comfort and acceptability of the work system to its users and other people affected by its use". For example, what the users think about the system's ease of use.

These 3 indicators are strongly correlated and can hardly be evaluated separately. An important lack of efficiency and effectiveness is likely to limit the system, and these indicators all have a direct impact on user satisfaction.

In addition, developers and experts can use ergonomic criteria to perform analysis of an interface. These criteria are expert tools based on large sets of individual recommendations. To facilitate the use of such criteria, Bastien and Scapin have produced a summary list of 18 criteria and have checked their effectiveness as a tool or a guideline for the evaluation of user interfaces [Bastien et al., 1993]:

1 Guidance

User Guidance refers to the means available to advise, orient, inform, instruct, and guide the users throughout their interactions with a computer (messages, alarms, labels, etc.), including from a lexical point of view.

1.1 Prompting (crit^{erion} 1): *refers to the means available in order to lead the users to making specific actions whether it be data entry or other tasks. This criterion also refers to all the means that help users to know the alternatives when several actions are possible depending on the contexts. Prompting also concerns status information, that is, information about the actual state or context of the system, as well as information concerning help facilities and their accessibility.*

1.2 Grouping / Distinction of items: *concerns the visual organisation of information items in relation to one another.*

1.2.1 Grouping / Distinction by location (crit^{erion} 2): *concerns the visual organisation of information items in relation to one another. This criterion also concerns the relative positioning of items within a class, i.e. within an area having a single functionality .*

1.2.2 Grouping / Distinction by format (crit^{erion} 3): *concerns more precisely graphical features (format, colour, etc.) that indicate whether or not items belong to a given class, or that indicate distinctions between different classes, or else distinctions between items of a given class.*

1.3 Immediate feedback (crit^{erion} 4): *concerns system responses to user actions. These actions may be simple keyed entries or more complex transactions such as stacked commands. In all cases, a fast response from the computer should be provided with information on the requested transaction and its result.*

1.4 Legibility (crit^{erion} 5): *concerns the lexical characteristics of the information presented on the screen that may hamper or facilitate the reading of this information. It does not concern feedback or error messages.*

2 Workload

Workload concerns all interface elements that play a role in the reduction of the users' perceptual or cognitive load, and in the increase of the dialogue efficiency.

2.1 Brevity: *corresponds to the goal of limiting the reading and input workload and the number of action steps.*

- 2.1.1 Concision (criterion 6): *concerns perceptual and cognitive workload for individual inputs or outputs. It does not concern feedback or error messages.*
- 2.1.2 Minimal actions (criterion 7): *concerns workload with respect to the number of actions necessary to accomplish a goal or a task. (The goal is to limit as much as possible the steps users must go through.)*
- 2.2 Information density (criterion 8): *concerns the users' workload from a perceptual and cognitive point of view with regard to the whole set of information presented to the users rather than each individual element or item.*
- 3 Explicit control
It concerns the users' workload from a perceptual and cognitive point of view with regard to the whole set of information presented to the users rather than each individual element or item.
 - 3.1 Explicit user action (criterion 9): *refers to the relationship between the computer processing and the actions of the users. (The computer must process only the actions requested by the users and only when requested to do so.)*
 - 3.2 User control (criterion 10): *refers to the fact that the users should always be in control of the system processing (e.g., interrupt, cancel, pause and continue).*
- 4 Adaptability
The adaptability of a system refers to its capacity to behave contextually and according to the user needs and preferences.
 - 4.1 Flexibility (criterion 11): *refers to the means available to the users to customise the interface in order to take into account their working strategies and/or their habits, and the task requirements.*
 - 4.2 User experience (criterion 12): *refers to the means available to take into account the level of user experience.*
- 5 Error management
Error Management refers to the means available to prevent or reduce errors and to recover from them when they occur.
 - 5.1 Error protection (criterion 13): *refers to the means available to detect and prevent data entry errors, command errors, or actions with destructive consequences.*
 - 5.2 Quality of error messages (criterion 14): *refers to the phrasing and the content of error messages (relevance, readability, and specificity about the nature of the errors) and the actions needed to correct them.*
 - 5.3 Error correction (criterion 15): *refers to the means available to the users to correct their errors.*
- 6 Consistency (criterion 16)
Consistency refers to the way interface design choices (codes, naming, formats, procedures, etc.) are maintained in similar contexts, and are different when applied to different contexts.
- 7 Significance of codes (criterion 17)
Significance of Codes qualifies the relationship between a term and/or a sign and its reference.
- 8 Compatibility (criterion 18)
Compatibility refers to the match between user characteristics (memory, perceptions, customs, skills, age, expectations, etc.) and task characteristics on the one hand, and the organisation of the output, input, and dialogue for a given application, on the other hand.

Testing methods

Usability evaluation can be user-based or expert-based. The method adopted for the evaluation of SPIRIT – and the only method detailed here - will be user-based. In ordinary user-based evaluation, the 3 usability indicators, effectiveness, efficiency and satisfaction, are measured with both a participatory evaluation and a questionnaire.

1) Participatory evaluation

To organise a participatory evaluation, it is important to identify the main user tasks and user groups. Representative users from each group are then selected. Experiments about the required number of users [Nielsen et Al, 1993] have shown that a selection of 8 representative users of each user group is generally sufficient to highlight the main issues.

A scenario, based on the identified user tasks is presented to the users and they are asked to work it out. The idea is to tell them what tasks to perform but not how to perform them.

During such a user session, interactions between users and the system are observed and encountered problems are noted down.

2) Questionnaire

At the end of the test, the users are asked to fill in a questionnaire. The questionnaire is the most frequently used tool for evaluating usability. Even if the questionnaire is generally called "satisfaction questionnaire", it also measures the system's effectiveness and efficiency.

Some well-known questionnaires (WAMMI³, QUIS⁴, SUMI⁵...) have been designed to measure the user satisfaction for any software system. They entail very general questions like "Is performing the task straightforward?" or "What do you think about organisation of the information on the screen?") grouped into general themes like ("Terminology and system information", "System capabilities", etc.). Most of the time, these are multiple choice questionnaire forms. An additional questionnaire focusing more precisely on the evaluated software functionalities is helpful to render the diagnostics more precise.

Participatory evaluation and questionnaires are two complementary means of measuring usability. However, examining the interactions between the user and the system is not an easy task. During user sessions, two other means are frequently used to help assessors to note and observe these interactions: the VPA method and the use of video recordings.

3) VPA (Verbal Protocol Analysis)

Verbal Protocol Analysis consists in asking participants to think aloud while performing their tasks [Ericsson et al., 1994]. This methodology is commonly used to address two goals:

- to study the mechanisms behind task-directed cognitive processes, and
- to improve tool design.

During a user-session, users are asked to verbalise what they are:

- doing (i.e. how they interact with the system)
- seeing (i.e. how they interpret aspects of the system)
- thinking (i.e. hypotheses they generate)

while performing tasks.

4) Video

Video recording is an added tool to the participatory evaluation. It is an easy way to record users' reactions toward the system. Moreover, use of video during a user-session is a way to avoid users' biased behaviours. Indeed, users' reactions can be modified with the presence of an assessor taking notes beside them, and thus the data measured may not be veridical.

Practically, two rooms are used so the users are separated from the assessors or evaluators. More than one evaluator may be involved, for example, ergonomics specialists, developers, interface designers, etc. Users are asked to perform tasks described in test scenarios. If the VPA method is used, the session is also recorded.

Two video cameras are normally used, one for recording users, their expression and actions and the other to record the progress as displayed on the monitor. Thus, the assessors have an opportunity for analysis during the user session, or after the user sessions have taken place.

³ www.wammi.com

⁴ www.cs.umd.edu/hcil/quis

⁵ sumi.ucc.ie

4. Evaluation tests for SPIRIT

4.1. Evaluation context

The SPIRIT evaluation methodology is grounded on the methods presented above.

However, SPIRIT will be evaluated at different stages of its development and thus the evaluation methodology will have to be flexible enough to fit both the interim prototype and the final prototype. Indeed, the functionalities will be different between both prototypes. Some of the tests discussed in this method can't be carried out for the evaluation of the interim prototype (e.g. the usability comparison between SPIRIT and operational search engines like Mirago or Local Google).

Moreover, it is not possible to have a performance comparison between SPIRIT and other similar systems. This is due to the fact that SPIRIT is indexing a specific collection of documents and it is not possible to plug other search engines to the same collection. Hence, the pooling approach used in TRECs which requires several search engines to constitute a reference document collection will not be used. The ISJ approach will therefore be adopted [Sanderson et al., 2004].

4.2. Stability test

The purpose of stability testing is to measure the influence of minor input variations (slight variations in user queries on a same search topic) on the SPIRIT results.

Methodology

Queries generated by users during the user requirements reassessment test (20 queries for each of the 3 topics treated in document WP7-D3-bis), or during a user evaluation test, will be used. They will be transposed to apply to the available test collection (which constrains the range of search places).

4.3. Global test for system usability and performance

Methodology

The evaluation method for SPIRIT combines both usability testing and performance testing. The test will be decomposed into two sessions, as shown in Figure 2.

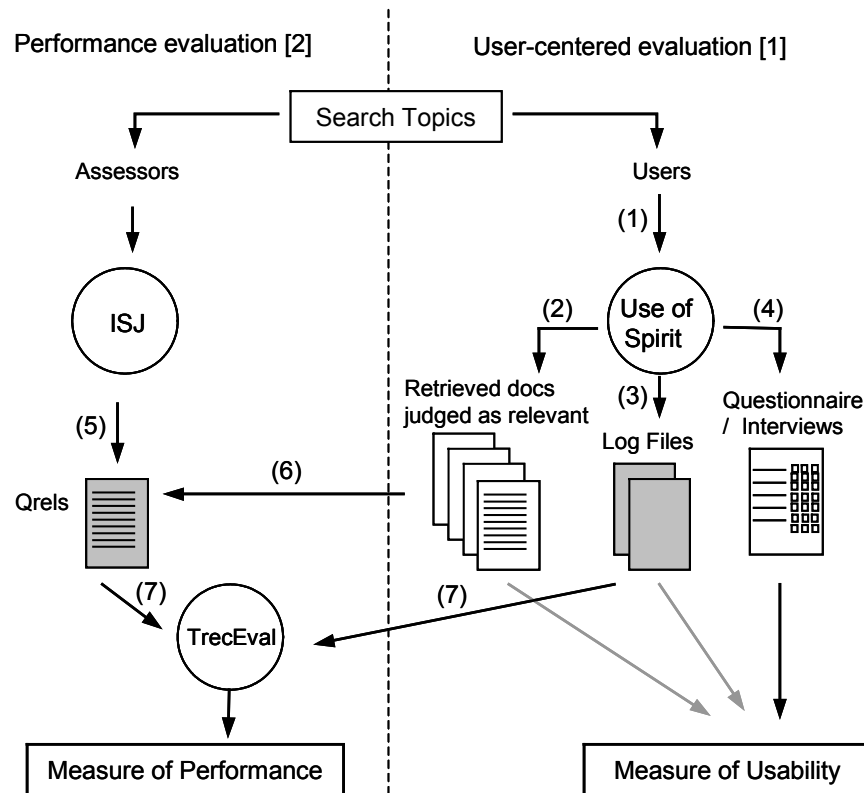


Figure 2: Global test for system usability and performance

User-centred evaluation

The first session will be user-centred and will measure usability [1]. Eight to ten users will be recruited according to the profiles specified in WP7-D3B [Bucher et al., 2004]: users who are *fairly familiar with information retrieval on the Web*, i.e. to use at least one Web search engine regularly. For the final evaluation a diversity in terms of profession, cultural background, and domains, such as travelling (including tourism), real estate, organising events and job hunting will be ensured.

However, for the interim prototype evaluation, diversity of profiles is not required as the interface is still very basic. The users with academic profiles will thus be recruited among assessors' relationships.

(1) Each user will be asked to use SPIRIT for several predefined search topics (all users will search for the same topics). These search topics will be illustrated to the users in search scenarios.

(2) During each search, the users will be asked to mark documents they consider as being relevant. This will result in a list of user-relevant documents.

(3) The interactions between users and SPIRIT will be recorded in log-files. (queries formulated by the users, list of web-sites retrieved by SPIRIT, time taken by SPIRIT for processing).

(4) At the end of the session, users will be asked to fill in questionnaires followed by an interview session. The following are the types of questionnaires to be used.

- A satisfaction questionnaire based on the generic QUIS questionnaire [Chin et al., 1988], and enhanced with questions based on ergonomic criteria inspired from the Bastien and Scapin summary list.
- A functionality-centred questionnaire. The questions here will depend on the available functionalities at the time of the test. These questions can also be asked orally.

(5) Finally, the assessor may interview the user to gather additional comments and explore ambiguous issues and problems pointed out by the users in the questionnaire.

Performance evaluation

The second session of the test will be used to evaluate performance [2]. The relevance judgments (Qrels) will be formed using the ISJ method (5). To take into account SPiRiT's spatial capabilities, ISJ searchers will be allowed to extend their queries not only to synonyms, but also to places contained in or near the searched place. The relevant documents retrieved during the user session will also be used to enrich the Qrels (6). Then, the log files will be directly used for the comparison between documents retrieved by SPiRiT during the user session and relevant judgments (7). This comparison will be automatically performed by TrecEval; the software used by TRECs to calculate both Precision and Recall. The use of TrecEval will be performed by the University of Sheffield.

Analysis of results

User-centred evaluation

The user session mostly relies on the results of the questionnaires, which should point out general and specific usability issues.

- Analysis of the QUIS-like satisfaction questionnaire

Even if questions are answered with quantitative grades (from 1 to 5), the questionnaire will lead to a rather qualitative analysis as follows.

The average grade, once calculated on all the questions, is translated into a global satisfaction grade. But this questionnaire also provides information concerning efficiency and effectiveness: indeed, even if each question refer to user satisfaction, many of them also refer to either efficiency (e.g. "Performing a task is straightforward") or effectiveness (e.g. "System reliability"). Then, the variation between those questions' grades and the average questionnaire grade is explored as an indicator of a specific usability issue.

- Analysis of the functionality-centred questionnaire and other data

The information on efficiency and effectiveness is completed using the list of user-relevant documents (how many documents have been retrieved) and the log files (information about the time spent on each search and the number of interactions).

The functionality questionnaire, which provides user feedback on SPiRiT specific functionalities, will enable us to identify the cause of the problems pointed out in the satisfaction questionnaire, and to propose directions for system improvement.

In addition, the interview is a way to easily understand and corroborate the results of the questionnaires. Indeed, for each user, the questionnaires are used as a basis for an interview where users are asked to comment on their answers, particularly those pointed out during the questionnaire session by a high variation in grade. The interview also gives the user the opportunity to highlight bad or good interface features that were not covered in the questionnaires.

Performance evaluation

The performance session provides quantitative data by comparing results of a SPiRiT search and relevant judgments (list of relevant documents for each topic). The TrecEval software calculates *Precision* and *Recall* indicators. These indicators are merely seen as an indication of SPiRiT's general performance.

4.4. Comparative test

As explained in section 3.1 it is not possible to compare SPIRIT with other search engines from a performance point of view.

However, a comparison will be made from a usability point of view, in order to measure the effect of the distinctive features of SPIRIT.

Methodology

Users are asked to execute several scenarios, eventually including a free search facility (to be decided later on, depending on the collection and available functionalities) on several search engines similar to SPIRIT (Local Google, Mirago, etc). Each search detailed in the scenarios uses geographical parameters and addresses SPIRIT functionalities.

After completing the tasks, users are asked to fill in a usability questionnaire. This questionnaire is composed of basic usability questions (e.g. "Do you find the results relevant?", "Is the search an easy task?", etc.). Users will answer the same questions for each search engine tested.

The questionnaire is then followed by an interview as described above.

Analysis of results

The comparative test session relies on the results of the comparative questionnaire, which aims at giving a user comparison of the different search engines. This questionnaire will only enable qualitative analysis as it relies on users' subjective point of view.

Each question of the questionnaire mainly focuses on one of the three usability criteria (effectiveness, efficiency and satisfaction). The grades obtained by each search engine will lead to a ranking of the different search engines tested, according to the three usability criteria.

The interview session conclusions, added to the questionnaire results, will be used to indicate what major ergonomics advantages and drawbacks SPIRIT has compared to other search engines. It will also show the effects of certain functionalities introduced in SPIRIT.

5. Conclusion

The purpose of this document is to provide a methodology for the evaluation of the SPIRIT interim and final prototypes. The methodology presented in this document integrates quantitative analysis that enables performance measurements and qualitative analysis that focuses on the usability of the system. The performance measurement is limited to the information retrieval functionality. Other evaluations, from a generic software point of view, are performed in WP1. The methodology proposed in this document will be used at different stages of SPIRIT's evolution. It must be used with flexibility, to adapt to the various stages of development.

This flexibility of the methodology is provided by a decomposition of the evaluation procedure into different elements that can be used independently. Moreover some elements may be specified depending on the available functionality to be evaluated. It is thus possible to use these elements to build two specific evaluation methods for the different SPIRIT prototypes: the interim prototype which has only a limited functionality, and the final prototype.

These specific evaluation methods will be described in future reports along with the results of each evaluation: deliverables D20 and D31, respectively.

Moreover, this methodology is likely to evolve after the evaluation of the interim prototype as this first evaluation will provide useful feedback concerning the relationship with users, the organisation of user sessions, etc. Thus, the conclusion of deliverable D20 will entail some clues for the improvement of the elements of the methodology presented here, and deliverable D31 will detail how the elements have been improved.

References

1. Amaryllis Web Site (1998) .www.upmf-grenoble.fr/adest/seminaire/inist
2. Bastien, J.-M.-C., Scapin, D.-L. (1993) "Ergonomic Criteria for the Evaluation of Human-Computer Interfaces" (version 2.1)
3. Bucher Benedicte et al. (2004) "User Requirements Specification Reassessment", WP7-D3-7101B
4. Chin, J. P., Diehl, V. A., Norman, K. L. (1988) "Development of a Tool Measuring Satisfaction of the Human-Computer Interface"
5. Cormack, G.V., Palmer, C.R., Clarke, C.L.A., (1998) "Efficient Construction of Large Test Collections", in Proceedings of 21st ACM SIGIR Conference, p. 282-289
6. Dave Finch et al. (2004) "Interim Pototype", WP1-D16-1101
7. Ergolab Web Site (2004) .www.ergolab.net
8. §Ericsson, K.A., Simon, H.A. (1994) "Protocol Analysis: Verbal Reports as Data revised edition", pub MIT Press, USA
9. ISO 9000-3:1997 Standard: Quality management and quality assurance standards -- Part 3: Guidelines for the application of ISO 9001:1994 to the development, supply, installation and maintenance of computer software.
10. ISO 9001:1994 Standard: Requirements for a Quality Management System Based.
11. ISO 9241 Standard: Ergonomic Requirements for Office Work with Visual Display Terminals.
12. Lespinasse, K., Kremer, P., Schribler, D., Schmitt, L. (1998) "Evaluation d'accès à l'Information Textuelle – Les experiences américaine (TREC) et française (AMARYLLIS)"
13. Loupy, C. and Bellot, P. (2000) "Evaluation of Document Retrieval Systems", LREC'2000, Satellite Workshop.
14. Nielsen, J. and Landauer, T.K. (1993) "A mathematical model of the finding of usability problems," Proceedings of ACM INTERCHI'93 Conference, Amsterdam, The Netherlands, pp. 206-213.
15. Quis questionnaire Web Site (2002) .www.cs.umd.edu/hcil/quis.
16. Sanderson, M. and Joho, H. (2004) "Forming Test Collections with no System Pooling", in Proceedings of the 27st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Sheffield, UK.
17. Sumi questionnaire Web Site (2002) .sumi.ucc.ie
18. Trec Web Site (2004) .<http://trec.nist.gov>
19. Usabilitynet Web Site (2003) .www.usabilitynet.org
20. Van Zwol, Roelof. (2002) "Modelling and Searching Web-Based Documents Collections", chap.3.4.1: Retrieval performance measure
21. Wammi questionnaire Web Site (2002) .www.wammi.com