# Benchmarking of filtering software and services – An Analysis Framework

*Definition of the Evaluation Criteria*

Version: Issue 1 Draft 2

| | | |
|---|---|---|
| ***Title:*** | Benchmarking of filtering software and services - Definition of the Evaluation Criteria. | |
| ***Abstract:*** | This document provides the definition of the evaluation criteria for a benchmarking process for filtering tools and services. | |
| ***Issue*** | Issue 1 Draft 2 | |
| ***Date:*** | February 16th, 2001 | |
| ***Authors:*** | Tom Jackson | Marco Riva, Francesco Puglisi |
| | Joint Research Centre | TXT e-Solutions |
| ***Distribution:*** | Unlimited | |

The role of the Joint Research Centre of the EC is to provide scientific support to the EU policy-making process by acting as a *reference centre* of science and technology for the EU. This report has been prepared by the Joint Research Centre in the frame of its institutional support programme to the EC DG Information Society. The opinions and views expressed in this report do not represent the official opinions and policies of the European Commission.

We invite readers of this report to send comments or suggestions to:

Tom Jackson
TP210, ISIS
Joint Research Centre
ISPRA (VA) 21020, ITALY
Tom.Jackson@jrc.it
Tel. +39 0332 786250

# 1 Introduction

This document is an interim deliverable of the study "Benchmarking of Filtering software and services – An Analysis Framework". It sets out a first draft of a set of evaluation criteria to be used in a benchmarking process for filtering software and services. The objectives of the benchmarking process are set out in the study Concepts Document[1], and a brief synopsis is provided below.

## 1.1 Study Context

The invention of the World Wide Web and of the graphics-enabled browser were the catalysts for a vast explosion of the availability of Internet content and of the number of Internet users world-wide, who now include an increasing proportion of families with children and schools which provide access for pupils. However, with the explosion of available information has come the problem of 'undesirable' or illegal data content, which is inappropriate for access by minors. One of the most widely publicised areas of undesirable material is the availability of so-called "adult" content, that is to say commercial pornography of various degrees, most of which is legal for adult consumption in the country of production. However, the Internet also serves as a distribution mechanism for other forms of undesirable or illegal content, such as racism, violence and paedophilia.

Surveys have shown that consumers (parents, schools, libraries, etc) view the unrestricted access to this material with great concern. Access control tools providing content filtering offer one approach to meeting concerns of consumers. These approaches have a number of advantages:

- They empower consumers to set the content standards depending upon their own set of values and on the age and maturity of the children involved;
- They provide a mechanism for consumer protection which can operate in parallel with legislative approaches, or, in the situations where enforcement of law fails, can operate independently as a self imposed monitoring scheme;

However, the performance of these tools, both in terms of usability and effectiveness, is variable, and often does not meet the requirements of the individual users.

# 2 Previous Work

Evaluations of filtering software have been carried by many diverse organisations, including government, academia, consumer organisations, freedom of speech groups, and in the traditional computer media[2]. Although these studies have been conducted for different social or scientific reason, and have taken differing technical approaches, they have consistently highlighted the

---

[1] Concepts Document: Benchmarking of Filtering Software and Services, T O Jackson, Oct 2000 http://e-filter.jrc.it

[2] For example:

**Access Prevention Techniques for Internet Content Filtering,** P Greenfield, P McCrea, S Ra, National (Australian) Office for the Information Economy, http://www.noie.gov.au/publications/NOIE/CSIROfinalreport.html

**Through the Net,** Which? Association - Filtering Test, May 2000. http://www.iwf.org.uk/safe/which/total.htm<x-html>

**Why Internet Content Rating and Selection does not work** (Kristian and Marit Köhntopp), http://www.koehntopp.de/kris/artikel/rating_does_not_work/

**Canada - Study on rating and filtering (Industry Canada)**. http://strategis.ic.gc.ca/SSG/it05082e.html

**Reviews of Internet Access Filtering Software (SuperKids)** http://www.superkids.com/aweb/pages/reviews/kidsafe/1/sw_sum1.shtml

**Access Denied:** The Impact of Filtering Software on the Lesbian and Gay Community, GLAAD, Dec 1997,

**Smut Filter Blocks All But Smut** (Wired) http://www.wired.com/news/technology/0,1282,36923,00.html

**Free speech advocate raises ire of filtering firms** (CNET News.com) http://news.cnet.com/news/0-1005-200-1567022.html

**Filtering the Internet: A Best Practices Model**, J M Balkin, B Noveck, K Roosevolt, Yale Law School, Yale University, New Haven (USA) http://stiftung.bertelsmann.de/internetcontent/english/download/Filtering.doc

weaknesses inherent in the existing software tools and services. Aside from political and cultural issues these criticisms are focused principally on the technical limitations of URL and keyword filtering. These limitations can be largely attributed to the problems of determining the context of the data being filtered. The tools have been shown to exhibit problems of adequately blocking harmful content and of also being prone to incorrectly blocking acceptable material. Technical problems have also been identified with content labelling approaches. These include questions as to who should do the labelling and how to ensure consistency of labelling across cultural boundaries.

These evaluation exercises have also highlighted the fact that there are no standard testing approaches for evaluating filtering tools, and the test results are largely derived from ad-hoc test methods.

A study that is particularly relevant to this benchmarking exercise is the work undertaken by the IDATE[3] project (a Preparatory Action for the Safer Internet Action Plan). This study evaluated the effectiveness of commercial software filtering tools and their suitability for European users. The study, which was largely based on pa an-European end-user survey, highlighted many of the current limitations of filtering tools and services from the perspective of the users. These included the following major issues:

- Difficult to install and configure for non-expert PC users;
- Filtering performance is erratic; harmful content is often not filtered and acceptable sites are inexplicably blocked;
- The diverse cultural and linguistic aspects of EU are not catered for by the tools (for example, offensive text is only recognised and filtered if written in the English language);

# 3 The Benchmark Proposal

In order to stimulate the development of improved filtering tools and services, and in order to improve awareness of the capabilities and limitations of filtering software, it has been proposed that a benchmarking process should be developed. The benchmark process will achieve these objectives though a number of distinct mechanisms:

- **Performance Goals:** By developing an independent and standardised test process that can rigorously evaluate the performance of filtering tools in all key aspects, including blocking effectiveness, functionality and usability, the relative strengths and weaknesses of filtering tools can be identified. As the demand for these tools increases (driven by increasing Internet access by minors and as a result of imposed local and national legislation) it is vital that the weaknesses in the current product base are identified and that both the consumer market and the producers are made aware of the limitations. It is anticipated that the identification of the weaknesses will spur technical development and refinement of the tools. The benchmark process will also provide a means of testing that a product meets a certain minimum required level of performance, especially in regard to blocking effectiveness.
- **Quality Assessment:** Published tests and feedback from user surveys (again, reference the IDATE report) has firmly established that quality issues are a major concern for users of filtering software. Consequently, a major aspect of the benchmark process will be quality assessment – that is, determining the degree to which the tools fit their intended purpose. Quality assessment will address many different functional and non-functional properties of the filtering tools, including user interface issues, installation and maintainability issues, and reliability and security issues. A range of measurement metrics will be defined that will be sufficiently generic to cover a broad class of filtering tools but which will also be sufficiently comprehensive to provide a detailed assessment of the relative quality of the tools. This assessment, as far as is practical and achievable will be based on quantitative measures rather than subjective qualitative measures.
- **Standardisation**: In order to better inform the end-users of the performance and functional characteristics of filtering software the benchmark study will provide a test

[3] IDATE, Prepact: Review of EU Third party filtering and rating software and services (Lot 3), Final Report, Vol. 1, Dec 1999, www.idate.org.

process that will become standardised across Europe. This will ensure that filtering tool evaluations can be carried out in a *systematic, reliable, repeatable* and *comprehensive* manner. Ensuring that the benchmark process is adopted as a standard approach, will require a number of key activities including; developing the test process within the context of known and established software evaluation standards (for example ISO), ensuring the acceptance and input of the producers of filtering software tools, soliciting the acceptance and input of third party stakeholders such as organisations involved in consumer protection or software evaluation.

The benchmarking study will also investigate the development of mechanisms for the partial automation of the testing process, most notably with regard to the blocking effectiveness. Comprehensive software evaluation is a time consuming and expensive process. To date, most filtering software trials have deployed relatively limited testing of the blocking performance (e.g. the tools are typically tested on less than 100 URL's) because the evaluations have used a manual test process. It is expected that the mechanism for evaluating both blocking and over-blocking performance (see section 4.1 for a definition of these terms) can be automated via web browser scripts, deploying standard programming methods such as JAVA or XML. These mechanisms will facilitate much broader test coverage for filtering performance.

# 4 The Benchmark Process

In the following discussion the structure of the benchmark process is explained. The benchmarking process has three principal components:

1. A set of evaluation criteria defining the measurements to be applied during the test process;
2. A test method that defines how the tests should be performed;
3. A data processing method that defines how the results of the tests should be processed.

## 4.1 Evaluation Criteria

The evaluation criteria define the measurement framework for the benchmarking process. Following a review of the literature on filtering testing and user–requirements it has been decided that the following assessment areas must be addressed by the benchmark:

- Blocking effectiveness
- Over-blocking sensitivity
- Security integrity
- Operational integrity
- Configurability
- Customisability
- Usability

These measurement factors encapsulate the broad range of functional and non-functional properties of the tools that are the major areas of concern for end-users. In the following sections, each of these parameters will be defined.

### 4.1.1 Blocking Effectiveness

The measure of blocking effectiveness we define as the **relative performance of the tool in blocking harmful[4] content.** That is, to what degree is the tool successful in preventing harmful internet content being displayed within a browser during an on-line Internet session. Clearly, this will be one of the most important measures applied to the filtering tools. If possible, the analysis should be quantitative rather than qualitative.

[4] See the EU report on Internet content rating for a description of what has been defined as 'harmful' content within the study. COM(96) 487 Final – Illegal and Harmful content on the Internet. Communication from the EC Green paper on the protection of Minors and Human Dignity in Audio-visual and information services.

### 4.1.2 Over-blocking Sensitivity

*Over-blocking* we define as the measure of the tools to tendency to incorrectly block access to legitimate web sites that do not contain harmful content. It should be noted that over-blocking has been consistently recognised as one of the major weaknesses of filtering software tools.

### 4.1.3 Security Integrity

Security Integrity is the term that will be used to describe the capability of the tool to prevent the filtering services being by-passed, by, for example mechanisms to defeat the password protection or techniques to alter or remove the filtering parameters.

### 4.1.4 Operational Integrity

Operational integrity we define as an assessment of a tool's stability in use, that is both its' reliability in use and its' effect on the reliability of other browser tools.

### 4.1.5 Configurability

Configurability is an assessment of a tool's flexibility in combining diverse filtering methods (for example, URL blocking and/or text matching) according to user requirements.

### 4.1.6 Customisability

Customisability we define as the degree to which the software filters can be customised or modified according to user preferences.

### 4.1.7 Usability

Usability we define as a measure of the ease of use of the tool, both from the perspective of installing and maintaining the tool and in regard to the ease with which it can be deployed during an on-line browsing session. Usability will also address issues such as the 'transparency' of the tool, cost effectiveness and traceability. Transparency we define as the accessibility of the filtering rules or lists that are used by the service or tool to block access to web sites. Cost effectiveness is a measure of the price to performance ratio of the filtering service or tool. We define traceability as the capability of a tool to monitor, log or trace the browsing activity of a user during an on-line session.

## 4.2 Test Method

The test method is in the process of development. It is highly dependent upon the definition of the measurement metrics. As such, the details of the test method will be elaborated in a subsequent report, after an exercise to validate the measurement framework, and after initial trials have been carried out on a sample of filtering tools.

## 4.3 Data Processing Method

In section 4.1 the set of evaluation factors were outlined. These measurement criteria will be subsequently broken down into subcategories of measurement metrics (Ref. section 7 of this document). An important feature of the benchmark process is the means by which the measurement metrics can be rationalised and interpreted, because the test will involve an extensive range of measurements. It has been decided that the benchmark process will use a technique called the FCM model (Factor-Criteria-Metrics). This method of data representation and visualisation provides a mechanism to structure the measurement data, to normalise and weight the measurement data, and to synthesise the display of measurement data across the metric set using kiviat diagrams (radar diagrams).

### 4.3.1 The FCM Model

The FCM (Factor – Criteria – Metrics) is a data representation model defined for the organisation and management of software metrics. This model was proposed and developed under REBOOT, an Esprit funded project. It is based on three main principles:

- ***Hierarchy***: each top-level attribute (for example reusability, portability, etc) represents a top-level node in a "quality tree" that facilitates the management of metrics complexity.
- ***Normalisation***: Each intermediate node has a representative value that is normalised in the range 0..1, with zero meaning worst case, one meaning optimal value and 0.5 representing the acceptance threshold. This normalisation allows a quick evaluation of a composite metric indicator (an FCM node).
- ***Synthesis***: Each FCM node is represented graphically with a kiviat diagram (radar diagram), permitting a quick overview of the value of the node (the bigger the area, the better the value) and of its components.

An example of a kiviat diagram is shown below, figure1:

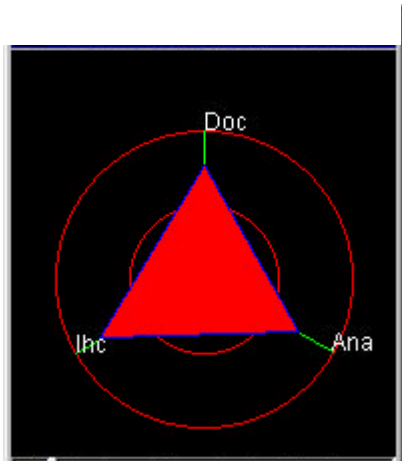



Figure 1: Kiviat (radar) diagram showing presentation of three measurement metrics.

# 5 Constraints and requirements for the benchmark process

In the following section the requirements and operating constraints for the benchmark process are discussed.

## 5.1 Defining a measure of Blocking Effectiveness

The most critical performance factor for filtering tools is the effectiveness of blocking harmful content. Consequently, the benchmark process will define a stringent test method for evaluating filtering performance. There are many factors that contribute to the overall filtering performance of a tool, and which must be considered during the definition of the test method. There are also benchmark process constraints that will define the structure and content of the data test set. The issues that will play a significant factor in the definition of the test method include the following:

- Filtering Tool configuration
- Data Test Set secrecy
- Data Test Set stability
- Data Test Set neutrality
- Cultural and linguistic diversity

### 5.1.1 Filtering Tool Configuration.

Commercially available filtering tools deploy a wide range of techniques for blocking internet content The most common of these are URL blocking lists or approved lists, text or key-word matching, and PICS label scheme. In order to ensure consistency of the benchmarking trials it is necessary to define a common **basepoint**, or **default-configuration**, for the tools. This is complicated by the fact that the tools do not contain the same combinations of available techniques, and may combine URL filtering lists with one or more of the other methods. Table 1

lists the comparative techniques and methods for cross section of common filtering tools. For the initial trials of the test process we have adopted the convention that the tools should be tested and configured at their **maximum capability**, that is, with all of the available techniques within a tool a specific tool activated for the trials. An alternative approach is to test each of the techniques within a tool separately, and in different sets of combinations. However, due to the complexity of this approach it will not be adopted for the initial development trials of the benchmark test.

| | CyberPatrol 4.0 | CYBERsitter 2000 | Safexplorer | NetNanny 4.0 | WE-BLOCKER | CYBER SENTINEL network |
|---|---|---|---|---|---|---|
| Stand-alone system | X | X | X | X | X | X |
| Protocol system | | X | | | | |
| **Filtering Technology** | | | | | | |
| Rating System | X | X | | | | |
| Keyword blocking | | X | X | X | X | X |
| Protocol blocking | | X | | X | | X |
| Host blocking | X | X | X | X | X | X |
| Characterised CR | | X | | | | |
| Time Restriction | X | X | | X | | X |
| **Host** | | | | | | |
| Client side | X | X | X | X | X | |
| Server side | X | X | | | | X |
| ISP's | X | X | | | | X |

Table 1: Comparison of filtering techniques offered by six common software filtering tools.

Having defined the default configuration, it is necessary to assess the other configuration parameters that influence the filtering performance of the tools. In this respect there are a number of issues which have to be clarified:

- **Which Lists**? The filtering performance of the tools is largely defined by the contents of the URL lists contained within the tool. Most tools contain two different types of lists which are typically described as **blocked URL lists** (the database of *prohibited* URL's stored within the tool, also known as 'black lists', 'No lists', 'Bad lists', etc) and **approved URL lists** (the database of *acceptable* sites, also known as 'white lists', 'Yes lists', 'Good lists'). In principle, the filtering performance in blocking *harmful* content should be perfect when using approved lists (dependent upon the subjective criteria used by the software suppliers to define *acceptable* content). However, many users have stated that deploying only approved URL lists (also known as a 'walled garden' approach) creates a browsing experience that it too restrictive for all but the youngest of children. Consequently, the test process will be conducted on the assumption that the tools should be configured for the trials so that they *deploy only blocked URL lists*, not approved lists. Tools will also only be tested with the lists supplied by the software providers. Additional user defined lists or third party sourced lists will not be used within the benchmark trials.
- **Which Keyword List?**: The filtering performance of the tools is also highly constrained by the keyword list that is deployed by the text-based matching algorithms. The keyword list is the list of 'prohibited' harmful words that a tool attempts to identify and block when examining a web page for content. Access to a web page is typically denied if one of the prohibited words on the list is identified on a page. Some tools come with preconfigured lists others are user definable. The test process will be conducted on the assumption that all of the tools will be configured with the same set of prohibited keywords.
- **Which Domain list?:** The argument identified for the configuration of keyword lists also applies to the problem of defining domain lists. Some of the filtering tools allow users to define domains (e.g. .com, .org) that can be blocked. The benchmark trials will be carried out on the assumption that domain blocking will not be activated during the evaluations.

- **What Age User**? The typical end-users of web browsers enabled with filtering software are children, either at home or in public amenities such as schools or libraries. However, the criteria for blocking harmful context is highly age sensitive – material that might be considered harmful or undesirable for young children may be perfectly acceptable for children of teenage years. The benchmark process must be independent of age factors. This may require different test sets configured to model the diverse age-related requirements of the users. Alternatively the test process may operate on the assumption of the highest required filtering scenario and test for blocking of content that would be considered harmful to the youngest age range. This latter approach will be adopted for the initial benchmark trials, due to time limitations on the study. However, it is acknowledged that end-users may find a series of trials organised into an arbitrary age hierarchy more satisfactory.

### 5.1.2 Data Test Set Neutrality

Testing of blocking performance of filtering tools and services has been typically carried out through 'hands-on' evaluation of the performance in accessing or blocking a random selection of unacceptable websites. For example, the free speech organisation Peacefire[5], which carried out a number of trials on filtering software, created a test composed from the first 1000 alphabetically listed **.com** URLs (although an arbitrary choice this list in fact contained URL's of over 300 web sites containing harmful content). The Which? Organisation[6] in their trials of May 2000, tested the software on a set of 23 sites considered to contain harmful content.

Testing of the software on an arbitrary list of URL's can provide some qualitative feel for the relative performance of the tools, but it is not a specific measure that can be used in a standardised benchmarking process. Also, the use of a small test set (relative to the number of available web sites) does not preclude the possibility of the test set being inadvertently biased in favour of the URL blocking lists of one or more of the tools. This could possibly lead to one or more tools having an apparent performance advantage over the others. Consequently, it will be necessary to define a data test set that has broad URL coverage to decrease the statistical possibility of inherent bias towards an individual tool.

### 5.1.3 Data Test Set Secrecy

If a database of URL's is used to test the filtering performance of the tools it is essential that the contents of the test remain undisclosed for the period that the benchmark is in force. If the contents of the test set are known then it creates the possibility for software suppliers to tailor the configuration of their software (e.g. by adding the test data to the default blocking lists) to perform well on the test set. Consequently, the list should only be distributed to trusted third parties or, alternatively it should be securely encrypted, and accessed during the benchmark trials via a deciphering script (for example, via a JAVA routine).

### 5.1.4 Data Test Set Stability

The data test set must remain consistent and stable for the duration that the benchmark is in force. This presents a problem for test sets that are based on the URL's of active web sites. The world-wide-web is a highly dynamic environment, and web sites are liable to change, close or move over short periods of time. The benchmark data test set must be independent of changes in web sites. There are two possible solutions to this problem. Either through the use of a '*virtual*' database of URL's, that remains static regardless of the actual content of the physical web site. Alternatively the benchmark process may be deployed with *dynamically* created test sets, that are valid only at the time of testing. The latter option would render it more difficult to standardise the benchmark.

---

[5] http://peacefire.org/censorware/Cyber_Patrol/first-1000-com-domains.html

[6] Through the Net – Which? Associaiton, Filtering tool Evaluation, May 2000, http://www.iwf.org.uk/safe/which/total.htm

### 5.1.5 Cultural and linguistic diversity.

The IDATE study reported that the broad range of commercially available software tools do not support the cultural and linguistic diversity of Europe. The tools are largely pre-configured to meet the requirements of English speaking users. Clearly, this poses problems for a benchmark process that is intended to be pan-European. However, this study will proceed on the assumption that the implementation of the benchmark process will be independent of language issues and cultural issues. The initial implementation will deploy testing only with respect to the English language. The method will be open to be adaptated for implementation to suit any cultural or language base.

## 5.2 Evaluation of usability related properties

A benchmark is typically used to measure the physical performance attributes of an item against a predetermined reference point or mark (for example, the established Business Winstone metric to assess the computational performance of a CPU). However, when dealing with software tools and services, it is typically the functional properties that are under assessment rather than physical properties. The evaluation criteria that have been defined cover a range of and functional and non-functional properties (e.g. performance, behaviour) of software filtering tools. The measurement metrics must be developed such that quantitative measurements can be made both for functional and non-functional properties. This may present difficulties in some assessment areas, most notably those associated with usability. Usability is a highly subjective software property, dependent upon diverse criteria such as user experience, user background, operating environment and personal preference. Consequently, usability issues will not focus on the operational and user interface aspects of the software, but rather on the more objective functional properties that allow the user to interact with the software tool and that determine the user's perception of a tool. Here we include issues such as traceability, transparency and cost–effectiveness.

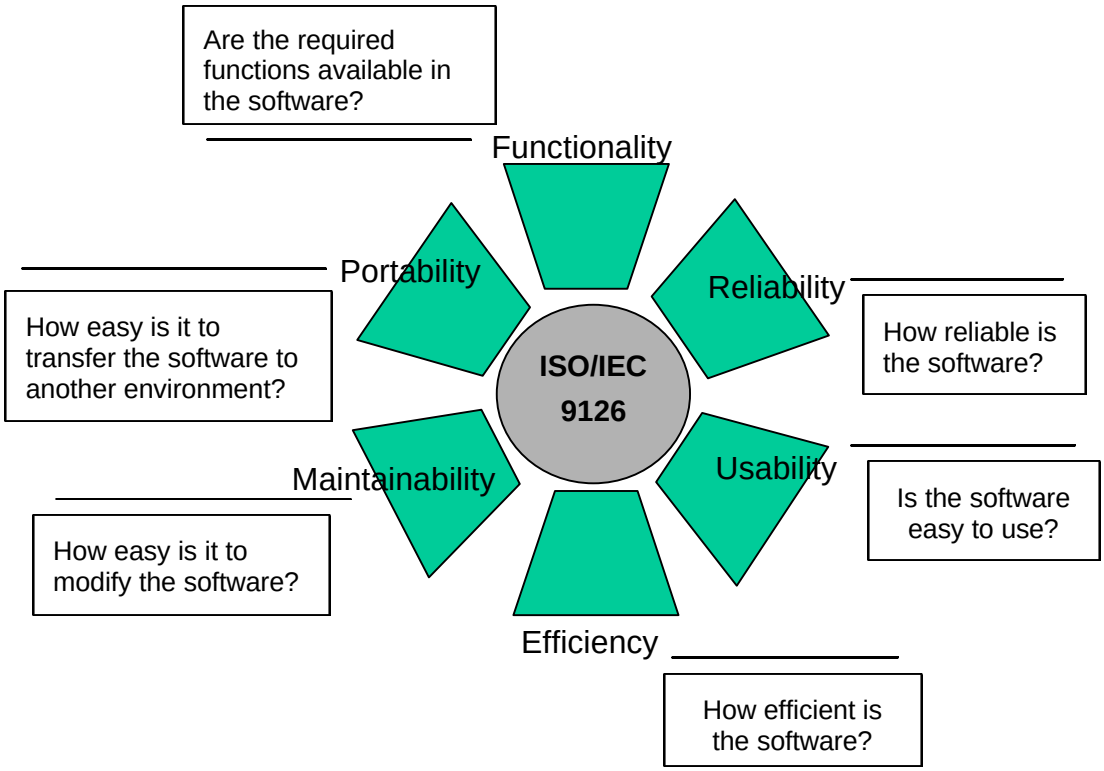
# 6 Filtering Software Quality Assessment.

The benchmark filtering process should become a standardised test method for Europe. This will ensure that European end-users have access to consistent evaluation information relating to the performance of filtering tools. To facilitate the standardisation of the benchmark process it has been decided that the test should fit within the framework of known and established software evaluation standards. To this end, the study has adopted the ISO 9126 standard for software quality assessment[7].

### 6.1.1 ISO 9126 Software Product Evaluation

The objective of the standard is to provide a framework for the evaluation of software quality. ISO/IEC 9126 does not provide detailed quality requirements for specific software products, but it defines a quality model which is intended to be applicable to all types of software applications. It defines six product quality characteristics, which are described in figure 2:

Figure 2: The Quality Factors defined by ISO/IEC 9126




[7] ISO/IEC 9216 – Information Technology – Software Product Evaluation.

ISO 9126 stipulates that the quality factors are divided into a number of quality subcharacteristics or criteria. These are described in table 2.

| Quality Factor | Quality Subcharacteristics |
|---|---|
| Functionality | Suitability<br>Accurateness<br>Interoperability<br>Compliance<br>Security |
| Reliability | Maturity<br>Fault Tolerance<br>Recoverability |
| Usability | Understandability<br>Learnability<br>Operability |
| Efficiency | Time Behaviour<br>Resource behaviour |
| Maintainability | Analysability<br>Changeability<br>Stability<br>Testability |
| Portability | Adaptability<br>Installability<br>Conformance<br>Replaceability |

Table 2: The Quality subcharacteristics defined by ISO/IEC 9126

## 6.2 The FCM Quality Model – ISO 9216 Derived

The quality framework, proposed under ISO 9126 is a generic model, for the analysis of diverse types of software product. However, for the purpose of the benchmarking study it is has been decided to further simply the model, and to introduce a new category that is specific to the benchmark requirements. The new framework, which will also be set within the FCM model discussed in section 4.3.1, is outlined in table 3. There are four principal quality factors, Functionality, Reliability, Usability, and Effectiveness. 'Effectiveness' is the category that has been introduced into the ISO9126 framework for the purpose of the study. ISO 9126 also provides annexes that recommend measurement metrics that are associated with each subcharacteristic. However, metrics have been uniquely defined for the benchmark process and are defined in section 7.

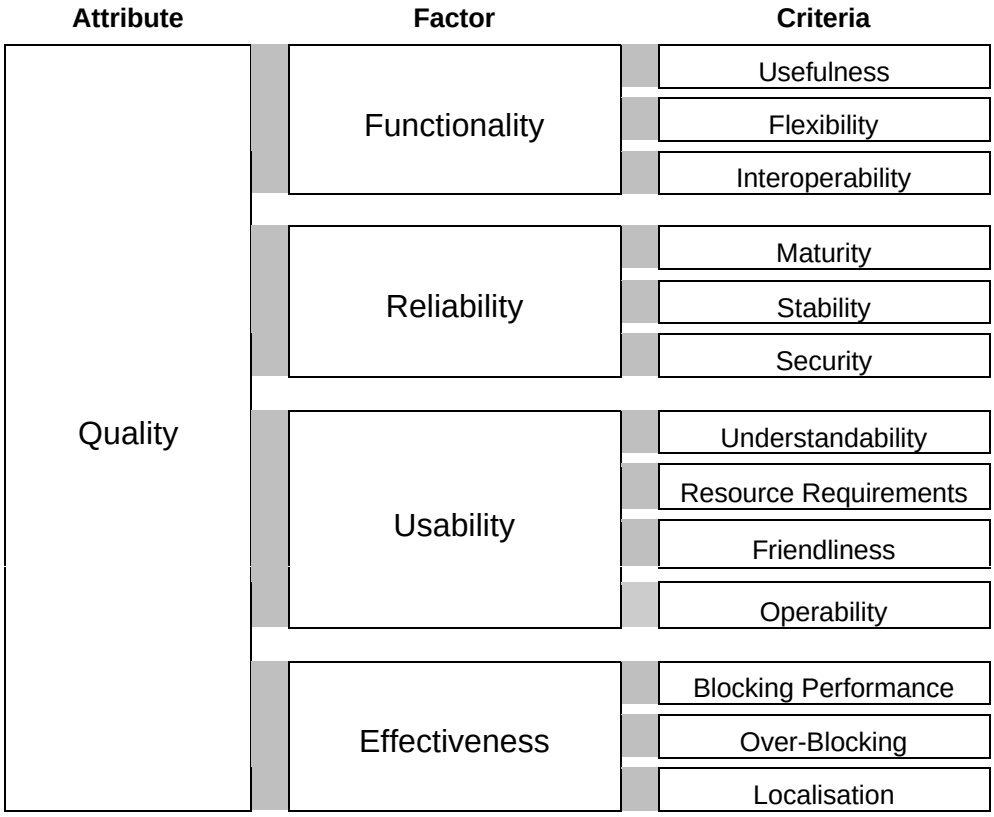
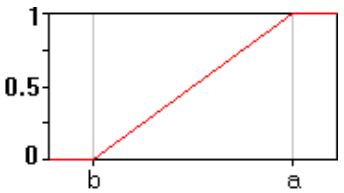



Table 3: FCM Model for Filtering tools Benchmarking

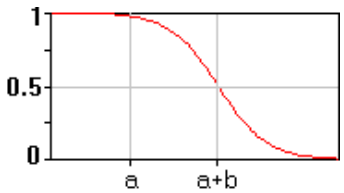
The defined Tree for QUALITY is an adaptation of the ISO9126 specification to the specific project's needs.

Each of the defined criteria is mapped to a number of metrics. Each of the single metric value is normalised in order to convert its value in the range 0..1. In the standard FCM model there are three different filters that can be associated with each metric: linear, exponential and gaussian. Each filter can be adjusted to perform the required conversion by properly setting the values for the parameters ***a*** and ***b*** (see the following picture). Further filters and metric weightings may be added during the test development phase.




**Linear filter**

f(a) = 1 ; f(b) = 0




**Exponential filter**

f(a) = 0.99 ; f(a+b) = 0.5


**Gaussian filter**

f(a)=1 ; f(a+b) = f(a-b) = 0.5

## 6.3 Software Quality: Factors and Criteria

In the following section we define the metrics that have been assigned for each quality subcharacteristic. Some of these criteria have been adopted directly from the ISO 9126

recommendations, others have been introduced to meet the needs of the benchmark evaluation process. For each set of quality criteria a set of measurement metrics have been defined, these are discussed in section 7.

### 6.3.1 F1: Functionality

*A Software application is held to be functional to the extent that the procedures it contains coincide with the functions required.*

In other words, the functionality refers to the compliance of the application with the user's expectations and requirements.

#### *6.3.1.1 F1C1: Usefulness*

Software attributes affecting the presence and the adequacy of all functions for a specific target.

#### *6.3.1.2 F1C2: Accuracy*

Software attributes concerning the generation of correct results or actions.

#### *6.3.1.3 F1C3: Interoperability*

Software attributes affecting the capacity to interact with specific systems.

### 6.3.2 F2: Reliability

The traditional definition of software reliability refers to an application's ability to maintain its specified performance whilst operating under specific conditions and for a specific period of time. For the purposes of the benchmark study we broaden the scope of this definition to include other performance variables such as security and the ability of the software to interact correctly with other, unspecified, software applications. Reliability in the context of filtering tools refers to the users' ability to operate the software within their host computer environment with a high degree of confidence that is going to fulfil all its functional requirements, whilst operating with stability and a sufficient level of security protection against misuse.

#### *6.3.2.1 F2C1: Maturity*

Software attributes affecting the frequency of failures due to software errors.

#### *6.3.2.2 F2C2: Stability*

Software attributes allowing the application to keep a specified performance level in case of software error.

#### *6.3.2.3 F2C3: Security*

Software attributes that bear upon the ability of the application to prevent unauthorised access to data or programs, either accidental or deliberate. Note that these criteria could equally well be placed under the *functionality* heading.

### 6.3.3 F3: Usability

The term *usability* refers to the attribute of 'user-friendliness' of a piece of software. User friendliness is something of a subjective concept, but in general terms ti refers to the ease with which a user can interact with a tool and gain operational access to totally exploit its functionality

The term 'users' typically refers to the direct operating agent of an interactive software application. Therefore users can be terminal operators, final users or in-direct users. The term can also refer to all those who are under the influence of or depend on the use of the application.

#### *6.3.3.1 F3C1: Understandability*

Software attributes affecting the effort made by the user in understanding the logical concepts of the software and its' functionality.

#### *6.3.3.2 F3C2: Resource Requirements*

Software attributes relating to the resources required to operate the software within the host computer environment.

#### *6.3.3.3 F3C3: Friendliness*

Software attributes affecting the effort made by the user in order to access to all the software features (e.g. number of clicks needed to perform an operation, or easy and intuitive access to all the features from a menu).

#### *6.3.3.4 F3C4: Operability*

Software attributes affecting the effort made by the user in operating and controlling the software.

### 6.3.4 F4 Effectiveness

This category is has been introduced into the ISO 9126 model, to specifically handle the analysis related to the filtering performance of the tools. There are three quality criteria.

#### *6.3.4.1 Blocking Performance*

Software attributes relating to the performance of the tool in correctly blocking harmful content.

#### *6.3.4.2 Over-blocking Performance*

Software attributes of the tool that lead to incorrect blocking of internet material that has no harmful content.

#### *6.3.4.3 Localisation*

Software attributes that relate to the ability of the tool to adapt to different cultural and linguistic operating environments.

# 7 Definition of the Measurement Metrics

In the following section we define a list of measurement metrics associated with each quality subcharacteristic. This is an initial proposal. It is expected that the metrics will be refined and, where necessary, extended, during the course of development of the evaluation process and through open-peer review.

## 7.1 Functionality

### 7.1.1 Usefulness Metrics

| | |
|---|---|
| URL blocking | Yes/No |
| Protocol blocking (TCP/IP) | Yes/No |
| Word blocking | Yes/No |
| ICQ chat blocking | Yes/No |
| newsgroup blocking | Yes/No |
| Email program blocking (Outlook Express, Opera,...) | Yes/No |
| Email attachments blocking | Yes/No |
| send/receive Email blocking | Yes/No |
| download blocking | Yes/No |
| Application blocking | Yes/No |
| File blocking | Yes/No |
| Control Panel access blocking | Yes/No |
| Rating system blocking (PICS) | Yes/No |
| Access time blocking | Yes/No |

### 7.1.2 Flexibility

| | | |
|---|---|---|
| URL Blocking Flexibility | Complete URL definition | Yes/No |
| | Partial URL definition (regular expression or other method) | Yes/No |
| Protocol Blocking Flexibility: | Ability to redirect to a specific port | Yes/No |
| Keyword Blocking Flexibility: | Complete word definition | Yes/No |
| | Partial word definition (regular expression or other method) | Yes/No |
| | Use of predefined words | Yes/No |
| Email Blocking Flexibility: | Ability to define blocking for specific accounts | Yes/No |
| Download Blocking Flexibility: | PDF, , , ( Napster ) | Yes/No |
| | ZIP | Yes/No |
| | DOC | Yes/No |
| | Multimedia Applications (e.g Naptster, Real Audio, MP3, MPEG, Quicktime) | Yes/No |
| Rating system Blocking Flexibility | SafeSurf or RSACi Activation | Yes/No |
| Access time Blocking Flexibility | Scheduling of allowed/forbidden time | Yes/No |
| | definition of maximum connection time | Yes/No |

### 7.1.3 Interoperability Metrics

| | | |
|---|---|---|
| Negative Impact on the operating system | | Yes/No |
| System interactions (co-operation with other filtering tools). | | Yes/No |

## 7.2 Usability

### 7.2.1 Understandability

| | | |
|---|---|---|
| Is there any notification about a blocking action | | Yes/No |
| If Yes, for each channel | | Yes/No |
| Is there an explanation of the reason for a blocking | | Yes/No |
| Transparency | User Access to “bad” URL list | Yes/No |
| | User access to “good” URL list | Yes/No |
| | User access to KEYWORDS list | Yes/No |
| | User access to PICS parameters | Yes/No |
| Is there a trace (log file) kept of the filter activity | | Yes/No |
| Is there a trace (log file) kept of the browser session activity | | Yes/No |
| Are log files analysable | | Yes/No |
| Are log files printable | | Yes/No |

### 7.2.2 Friendliness

| | | |
|---|---|---|
| Is there a default installation configuration | | Yes/No |
| If Yes, For each channel | | Yes/No |
| Is there a configuration wizard | | Yes/No |
| Is there a User manual (paper-online) | | Yes/No |
| Completeness of user documentation | Poor | Yes/No |
| | Good | Yes/No |
| | Excellent | Yes/No |
| Understandability of the user documentation | Poor | Yes/No |
| | Good | Yes/No |
| | Excellent | Yes/No |
| Is there a multimedia tutorial | | Yes/No |
| Is there a contextual help | | Yes/No |
| Is there an online F.A.Q. list | | Yes/No |
| Is there customer support? (Mail – Telephone – freephone etc) | | Yes/No |
| Is there local technical support (US/European/ single country…) | | Yes/No |
| Number of “click” needed to access to the functions (add new URL to list, download new URL list, etc…). | | Yes/No |
| Understandability of the user interface | Poor | Yes/No |
| | Good | Yes/No |
| | Excellent | Yes/No |
| Is there a multimedia tutorial | | Yes/No |
| Accessibility to functions through menu, icons, keyboard shortcuts, etc | | Yes/No |
| Effort required to learn how to operate the application | Easy | Yes/No |
| | Moderate | Yes/No |
| | Hard | Yes/No |

### 7.2.3 Resource Requirements

| | | |
|---|---|---|
| Static disk usage (installation) | | n |
| Dynamic disk usage (runtime) | | n |
| Memory usage (runtime) | | n |
| Processor Requirement | | type |
| Observable difference in browser response time with filtering active | | Yes/No |
| | If yes: < 1 sec | Yes/No |
| | < 10 sec | Yes/No |
| | >10 sec | Yes/No |
| Cost | Trial version available | Yes/No |
| | Low (0-20$) | Yes/No |
| | Medium (20-100$) | Yes/No |
| | High (>100$) | Yes/No |
| Cost of List updates (monthly) | Low (0-5$) | Yes/No |
| | Medium (0-20$) | Yes/No |
| | High (> 20$) | Yes/No |

### 7.2.4 Operability

#### *7.2.4.1 Configuration*

| | | |
|---|---|---|
| Blocking list updates | Automatic | Yes/No |
| | Manual | Yes/No |
| User defined filtering configuration (ie which combination of techniques) | | Yes/No |
| User definable URL blocking lists | Remove | Yes/No |
| | Add | Yes/No |
| | Import/Export | Yes/No |
| User definable keyword lists | Remove | Yes/No |
| | Add | Yes/No |
| | Import/Export | Yes/No |
| Define access privileges for each URL list | | Yes/No |
| Predefined URL Grouping categorisation | | Yes/No |
| Personalised lists definition | | Yes/No |
| User defined URL Grouping categorisation | | Yes/No |
| Create User profiles | | Yes/No |
| Define User privileges | | Yes/No |
| Define User access-time schedule | | Yes/No |
| User definable interaction with other applications | | Yes/No |
| Blocking of Port Numbers | | Yes/No |

#### *7.2.4.2 Installability*

| | | |
|---|---|---|
| Installation CD ROM | | Yes/No |
| Internet download | | Yes/No |
| Auto installation available | | Yes/No |
| Installation Wizard for User/System parameters | | Yes/No |
| Number of installation steps | | n |
| On-line help for installation | | Yes/No |
| Uninstall option | | Yes/No |
| Can user easily re-try setup installation of software? | | Yes/No |
| Can user or maintainer easily install software to operation environment? | | Yes/No |
| Easiness of manual install operation | | |
| | [very easy] only user's watching except just start install or setup functions; | Yes/No |
| | [easy] only user's answering to question from install or setup functions; | Yes/No |
| | [not so easy] user's looking up parameters from tables or fill-in-boxes to be changed and setting them; | Yes/No |
| | [complicated] user's seeking parameter files, looking up parameters from files to be changed and writing them. | Yes/No |

## 7.3 Reliability

### 7.3.1 Stability

| | | |
|---|---|---|
| Possible to install tool with other filtering applications | | Yes/No |
| Interoperability and compliance with other applications (non-browser) | | Yes/No |
| Meantime between errors | | time |
| Browsers supported | Netscape | Yes/No |
| | MS Internet Explorer | Yes/No |
| | Opera | Yes/No |
| | Others | Yes/No |
| Uninstall facility | | Yes/No |

### 7.3.2 Maturity

| | | |
|---|---|---|
| OS supported | Win 95 | Yes/No |
| | Win 98 | Yes/No |
| | Win 2000/me | Yes/No |
| | Win NT | Yes/No |
| | Unix/Linux | Yes/No |
| | Macintosh OS | Yes/No |
| | Solaris | Yes/No |
| Browsers supported | Netscape | Yes/No |
| | MS Internet Explorer | Yes/No |
| | Opera | Yes/No |
| | Others | Yes/No |
| Length of time tool has been commercially available | | time |

| | | |
|---|---|---|
| Filtering Algorithms | Standard | Yes/No |
| | Advanced | Yes/No |
| | State of the Art (e.g. image analysis) | Yes/No |

### 7.3.3 Security

| | | |
|---|---|---|
| Administrator's access password protected | | Yes/No |
| Support for hardware security devices (e.g smartcards) | | Yes/No |
| How easily can the password be disabled | Basic PC knowledge [easy] | Yes/No |
| | Normal PC knowledge [moderate] | Yes/No |
| | Expert PC knowledge [hard] | Yes/No |
| How easily can the filtering software can be made disabled | Basic PC knowledge [easy] | Yes/No |
| | Normal PC knowledge [moderate] | Yes/No |
| | Expert PC knowledge [hard] | Yes/No |
| How easily can the filtering software be removed | Basic PC knowledge [easy] | Yes/No |
| | Normal PC knowledge [moderate] | Yes/No |
| | Expert PC knowledge [hard] | Yes/No |

## 7.4 Effectiveness

The effectiveness metrics are performance measures. Consequently, they do not have simple measurement metrics, as defined for the above quality criteria. The metrics will be derived from active performance evaluations, as discussed within section 5 of this report. Detailed discussions of the effectiveness metrics will be issued in a subsequent report that defines the test method.

# 8 Conclusions

This report has presented a first draft of a set of evaluation criteria and quality factors for the analysis and benchmarking of filtering tools and services. The document has also discussed requirements relating to the design of the analysis framework and performance metrics. Additionally it has described the context of the evaluation process within an internationally recognised ISO standard for software product evaluation. Comments and feedback in regard to the proposed evaluation framework are welcomed. This can be achieved by direct contact with the author, or via the web site established to support the study.

http://efilter.jrc.it

This site hosts an interactive discussion forum as well as acting as an on-line repository for documents relating to the study.