

Early Results from Automatic Accessibility Benchmarking of Public European Web Sites from the European Internet Accessibility Observatory (EIAO)

Nils Ulltveit-Moe¹, Mikael Snaprud¹, Annika Nietzio², Morten Goodwin Olsen¹, and Christian Thomsen³

1) Faculty of Engineering, Agder University College, NO-4876, Norway

Email: nils.ulltveit-moe@hia.no, mikael.snaprud@hia.no, morten.g.olsen@hia.no

2) Forschungsinstitut Technologie und Behinderung (FTB) der Evangelischen Stiftung Volmarstein, DE-58300 Wetter (Ruhr), Germany, Email: eiao@ftb-volmarstein.de

3) Dept. of Computer Science, Aalborg University, DK-9000 Aalborg, Denmark, Email: chr@cs.aau.dk

ABSTRACT

Benchmarking of web accessibility is performed throughout Europe, to assess and raise awareness of web accessibility. The evaluation is often based on manual assessments with a high cost and with long intervals. The Web Content Accessibility Guidelines from W3C/WAI are the basis of most evaluations. Although the same guidelines are used, a range of different evaluation methodologies and scoring schemes are deployed across the member states. This makes it hard to compare and evaluate the web accessibility between the different countries within Europe. To improve this situation, the Unified Web Evaluation Methodology (UWEM) [1] is being developed as a joint effort between the projects in the Web Accessibility Benchmarking (WAB) cluster[2] to provide a unified way to evaluate web accessibility and ensure that both automatic large scale and manual assessment are possible within the same scheme. The European Internet Accessibility Observatory (EIAO)[3] is currently developing a prototype Observatory to perform large scale automatic accessibility benchmarking compatible with UWEM, with a plan to evaluate and present results from 10 000 public European web sites monthly. We will in this paper present some initial results, including an analysis of these results, from an early prototype of the Observatory. The results will cover five carefully selected web sites from each EU and EFTA country. These sites are meant to be representative of each country and include the prime minister site, national bank, national library and selected sites among federal organisations such as president/monarchy, ministries, parliaments and national statistics agency. The evaluation results will be presented as scores according to UWEM, based on aggregated barrier probability indicators and also the number of detected barriers for each site and country. Analysis of the numbers shows that it is possible to use the UWEM approach to sample and aggregate indicators from a random set of web pages using automatic assessment until a required error margin has been achieved. In many cases, the variance between samples for a web site is low and thus allows a representative indicator to be based on small fraction of the total number of web pages of the web site.

1 INTRODUCTION

Web accessibility benchmarking is carried out in many European countries to assess the

accessibility status and to increase the general awareness. Different countries have different benchmarking methods and carry out the assessment with different frequency. Even if the evaluations are based on the same guidelines, accessibility evaluations are in practice often carried out in different ways. This is preventing international comparisons and systematic monitoring of this basic requirement for a democratic development of the information society.

A tool enabling frequent and automatic evaluations at a low cost will allow policy makers to monitor the development more closely to identify good practices, allow for regional comparisons, and assess the impact of policy measures.

The European Internet Accessibility Observatory (EIAO) [3] is a project designed to provide a prototype of such a tool, namely an automatic large scale web evaluation service producing data on the accessibility status and development, focusing on public content. The final version of the prototype Observatory will publish monthly updated measurements from 10.000 web sites. The results will be available online from a data warehouse to support flexible analysis, provide a basis for policy-making, research and actions to improve the accessibility to Web content.

The Web Content Accessibility Guidelines (WCAG) [4], developed by W3C [5] Web Accessibility Initiative [6] have been adopted for public content by many national governments. Based on those guidelines three projects including 23 partners in the WAB cluster [2] have developed the Unified Web Evaluation Methodology (UWEM) [1]. The objective of UWEM is to provide means of ensuring that large scale monitoring and local evaluation are compatible and coherent among themselves and with the WCAG. While small scale evaluations can be performed manually, large scale evaluations require support of automated tools.

In this paper we present some initial results, including an early analysis of these results, from a prototype of the Observatory. The results cover up to five carefully selected web sites from each EU and EFTA country. These sites are representative of each country as they include the prime minister site, national bank, national library, and selected sites among federal organisations such as president/monarchy, ministries, parliaments, and national statistics agency. The evaluation results are presented as scores according to UWEM, based on aggregated barrier probability indicators and also the number of detected barriers for each site and country.

2 THE EIAO OBSERVATORY

UWEM specifies which checkpoints can be tested automatically, how the assessment results are sampled and aggregated and outlines how the results can be presented.

EIAO has implemented a prototype Observatory that performs large scale automatic assessments of web sites. The selected Open Source approach is essential to make sure that measurements of access to information – a prerequisite for any sound e-Government benchmarking – are carried out in a transparent and democratic way. Thus, allowing inspection of how the measurements are actually implemented, and encouraging a collaborative approach to improve them.

2.1 Architecture

The main elements of the first release of the Observatory are a crawler, a data warehouse, a set of Web Accessibility Metrics and a web user interface. Further description of how the Observatory works is given in the remainder of this section. For a more in-depth description of the system architecture, please refer to [7].

The key component in the EIAO Observatory architecture is the web crawler, which is based on the

Open Source web crawler HarvestMan [8]. HarvestMan has been chosen because it is a good and mature web crawler that was easy to adapt to the system. Moreover, the HarvestMan developers were interested in actively supporting the EIAO project. HarvestMan has been integrated with the other components of the Observatory and is one of the key component of the EIAO Software.

A set of Web Accessibility Metrics (WAMs), which are formal rules specifying how to make a statement about accessibility barrier indicators of a given web resource, has been specified based on the tests described in UWEM 0.5, checking for deviations from WCAG 1.0 checkpoints. The first set of WAMs is based on Relaxed [9] and Schematron [10]. More complex rules may be implemented in Schematron (or in some other language) in a later release. The WAMs for the next release of EIAO are based on UWEM 1.0. In addition to fully automatable tests for HTML there will also be tests assessing the accessibility of CSS.

2.2 Sampling strategy

The core activity of the Observatory includes crawling a selected number of web sites and performing automatic accessibility measurements on these sites. In order to do this, we have chosen to acquire samples from each site by modifying the HarvestMan web crawler to support sampling. It is a prerequisite for the validity of statistical analyses of the results that the set of pages chosen is random.

Because of this, the sampling strategy of the Observatory is based on the near uniform random sampling strategy presented in [11]. This means that no part of the web site should be favoured. Thus it is important to treat all link pathways within a web site equally. Read more on how the number of sample can be limited in section 3.3.

In short the following is the basic functionality of the near uniform random sampling during the sampling of one web site.

1. An already known seed URL is chosen, among all known URLs of the site¹, as a starting point. This could be a URL the main page of the site or any other known page of the site.
2. The chosen page will be crawled and all internal links (links within the same site) extracted.
3. Each of the extracted links in the page are chosen with a predefined probability d and thus not chosen with a probability $1-d$.
4. If no links are chosen, the random walk will end at the current page. However, if any link is chosen the algorithm will jump to 2.

A crucial point of this algorithm is the predefined d -values in which URLs to crawl are chosen. Uniformly distributing the d -values will result in favouring pages with a large number of links pointing into them, and because of this the sample strategy will not be uniform. In contrast, the links should be weighted according to how large part of the web site is reachable by following each link. The above algorithm can easily be followed if the structure of the web sites crawled are known.

However, this is most often not the case when crawling unknown web sites. As shown in [11] the number of paths available by following a link can be used as an approximation of this complexity. Because of this the d -values should be chosen based on the paths available by following a link. The d -values are incrementally updated during the crawl.

¹ Note that all new URLs found during random walk crawling are added to the repository of URLs, so that the entropy of URLs increases in subsequent crawls, as larger parts of the web site is being explored by the random walk. Currently we have on average 800 URLs to choose between for a web site when performing the assessments.

2.3 Aggregation Model

Figure 1 gives an overview of the web accessibility evaluation process implemented by EIAO. In the first stage a number of accessibility tests are performed. In release one of EIAO these tests correspond to the subset of UWEM 0.5 tests [1] that can be performed automatically. The tests identify the potential accessibility barriers (b_0, b_1, \dots) in page p . The second stage performs aggregation of the individual test results into one comprehensive number for a web page (F_p). The calculation is based on a probabilistic model. It takes into account the number of fail results of the test for each barrier type b on page p , which is denoted by R_{pb} in the figure. Additionally, there are parameters F_b that model the severity of each barrier type. The results reported in this paper assume the same severity value for all barriers. The result of the second stage (F_p) can be interpreted as the *accessibility barrier probability* for page p , i.e. the probability that the web page constitutes an accessibility barrier for a user with a disability.

To present the results to the public the third stage of web accessibility evaluation needs to provide means of interpreting the results. This includes a statistical analysis of the findings. EIAO calculates F_s , the mean barrier probability for each site s . As described in section 2.2 EIAO does not assess all pages from a web site but only a sampled subset. The margin of error (at 95% confidence) provides information about the quality of the sample.

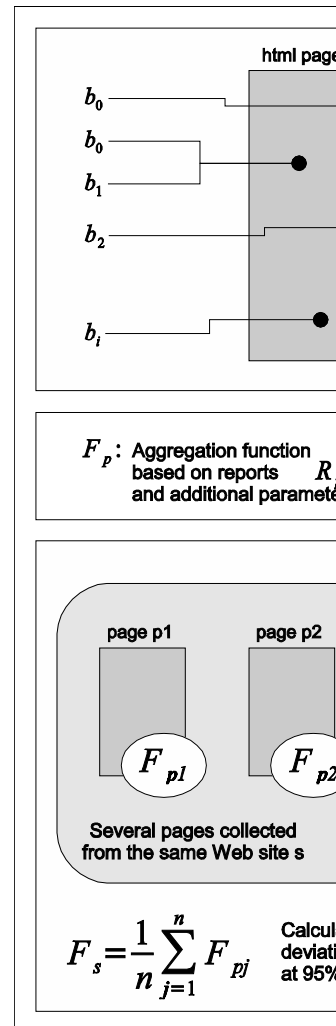


Figure 1:
Evaluation
stages

From the single web site results further aggregation steps are possible. These include the calculation of average barrier probabilities for groups of web sites.

2.4 Data Warehouse

A data warehouse named EIAO DW is used to store the large amounts of collected accessibility data. To support easy and flexible analysis, the data is stored in a so-called multidimensional schema as described in [12]. The data warehouse is implemented in PostgreSQL which was chosen since it is a very mature and extensible well-performing Open Source Data Base Management System with useful features such as table partitioning and bit-mapped indexes. Materialized view support which would be very beneficial for a project like the EIAO DW is still not present in PostgreSQL, but some projects are planning to add it.

The measurements stored in the data warehouse will be made available on the web. The data warehouse will contain the data from monthly measurements. In this way, trend analysis can be carried out, for example to assess the impact of some change in national implemented public procurement policy or accessibility legislation.

The results will be barrier probabilities that are possible to present with regards to a single web sites, NUTS regions (Nomenclature of territorial units for statistics) [13] and different categories. For an example of a report of such aggregations see section 3.1. These reports will in the final release of the Observatory be available for a given point-in-time as well as for a time period. In this way, comparisons of the status and development in e.g. local municipalities in different countries or regions will be supported.

Subsequent evaluations of these reports will guide future development of the functionality related to the detailed reporting. Based on feedback from end users (e.g. policy makers, associations of disabled people etc.), the tools for collecting, assessing and disseminating data will be continuously improved. User testing will allow improving the relevance of the automatically collected data.

2.5 Reliability and Quality Assurance

The usefulness of the Observatory depends heavily on the validity and reliability of the presented data. Unreliable data might lead to false interpretations and generalisations that generate wrong conclusions about web accessibility of a certain web site or group of web sties.

Therefore, we have taken precautions to minimise the inexactness during design, implementation and running of the Observatory.

The EIAO WAMs (Web Accessibility Metrics) are an implementation of the fully automatable UWEM tests². To make sure that the WAMs produce correct results, a conformance testing framework was developed. For each test at least one PASS and FAIL test case is provided. It consists of an HTML file and an EARL [14] report describing the expected test outcome. The result of the assessment is then compared to the expected result. The use of EARL allows on automatic comparison.

The conformance testing framework is the most basic mechanism for reliability assurance. It is run regularly as a regression test after changes concerning the WAMs to ensure that the outcomes of the assessments performed in different test runs on the same test suite are correct.

The URL directory is used to store all URLs identified for the websites being assessed, and in addition state data for the random walk algorithm. (See section 2.2 for description of random walk and the state information needed). To ensure that the sampled data have good enough entropy, we require that the random walk algorithm has identified at least 10 URLs after a crawl. Web sites with less than 10 URLs need to be investigated manually, to identify if all available URLs have been identified, and if not, the reason why these URLs were not identified.

The statistical reliability of the results presented in the EIAO data warehouse depends on the URL set. For reports about groups of web sites it is crucial that the URL collection is representative for the assessed categories.

The initial URL collection has been carried out by a group of experts. EIAO partners from several countries were included to quality assure the results of the sites. Since this collection is used as a seed for an automatic extension it has been carefully checked to contain only relevant URLs for the

² Note that this paper describes the first EIAO release which is based on UWEM 0.5 tests. UWEM 1.0 is available since July 2006 and will be the basis for the next EIAO release

areas that were investigated.

3 RESULTS OF THE ACCESSIBILITY ASSESSMENTS

As described above, an essential part of the Observatory includes calculating the *accessibility barrier probability indicator* for each sample. Table 2 shows the average score for each sample for all evaluated web sites in the June crawl using $F_b=0.05$ for all barrier indicators b^3 . For reference purposes, the UWEM score scale is presented in table 1. The scores per web page are aggregated into a complete accessibility barrier indicator score for the entire site, shown as UWEM score. The UWEM score can further be aggregated for a complete region, country or a sector. An example of such an aggregation can be seen in figure 1.

The results cover up to five carefully selected web sites from each EU and EFTA country. These sites are meant to be representative of each country and include the prime minister site, national bank, national library, and selected sites among federal organisations such as president/monarchy, ministries, parliaments and national statistics agency.

Furthermore, the average barrier probability indicator for a web site is presented under F_s , while the error margin from the average result for each site can be seen in the column ERRM95%CI. In the table all sites with an error margin of more than 0.07, or less than 10 URLs or samples identified have been excluded as unreliable crawls. We have also excluded web sites where the change in barrier probability was significantly larger than the error margin between the June and July crawl, since we need to do further analysis in order to reach a conclusion for these web sites. Please note that the score A is almost impossible to achieve by automatic assessment, since it requires no barrier indicators to be found. The Observatory will in most cases find possible barrier indicators, however human judgment is needed to verify if the barrier indicator constitutes a real barrier. Also note that the automatic testing covers only a subset of all WCAG checkpoints because many checks are not possible to automatise. This means that the Observatory is able to identify some web sites with poor accessibility; but manual evaluation is needed to evaluate a more complete set of tests.

Table 1: UWEM Score Scale

Score	Barrier Probability F_s
A	$F_s = 0\%$
B	$0\% < F_s < 25\%$
C	$25\% < F_s < 50\%$
D	$50\% < F_s < 75\%$
E	$75\% < F_s < 100\%$
n/a	Not available

³ We are still doing research on F_b values, and the recommended values may change in the future. $F_b=0.05$ for all barriers b can be interpreted like 5% of occurrences of a given barrier indicator b constitute a real barrier for the user.



Table 2: Barrier Probability indicator of some European Governmental Web Sites (assessed in June 2006)

3.1 Graphical User Interface (GUI)

The aggregated UWEM scores can be presented for each site as in table 2 above, or further aggregated into scores for e.g. European countries. An example of such an aggregation can be seen in figure 2 from the EIAO GUI presenting results from June 2006.



Figure 2: Barrier probability of European Web Sites

3.2 Discussion of results

The scores presented in figure 2 are based on the assumption that $F_b=0.05$ is a reasonable choice of barrier severity. This value was chosen to have the UWEM aggregation algorithm within its operating range for a choice of real web sites. It is ongoing research to improve the aggregation model in UWEM, and the choice of F_b . For more details about the various aggregation models please refer to [15]. The analysis in this paper studies whether it is possible to achieve repeatable and coherent results by acquiring random sampled pages with the random walk strategy outlined above and the UWEM aggregation algorithm.

It is noticeable in table 2 that the average barrier probability for each sample (F_s) varies for each web site crawled. It is also noticeable that the error margin varies for each site. This means that in order to reach an equal error margin of the calculated barrier probabilities the number of samples drawn could be adjusted in contrast to only sampling a predefined 100 samples. Assuming a normal distribution the error margin within a 95% confidence interval can be calculated as $1.96 \sigma \sqrt{N}$, where σ is the variance and N is the number of samples. This means that the error margin will improve by the square of the number of samples, and the error margin is directly proportional to the variance. In order to reach an equal error margin for all web sites, the number of sampled pages drawn from some sites may be reduced, while the number of samples for other sites will need to be increased.

A direct conclusion of this observation is that varying the number of samples drawn based on error margin, otherwise known as adaptive sampling, would be most beneficial. This might increase both the reliability of the calculated barrier probability indicators, as the error margin would be close to equal for all sites, and decrease the number of samples needed. The question remains if it is viable to perform random walk sampling to an error margin that could be tolerable within the scorecards used by UWEM.

3.3 Adaptive sampling experiment

In order to determine how beneficial such adaptive sampling could be, we did a simulation of how many samples this would require, to achieve a 5% error margin for the site average barrier indicator with 95% confidence interval using the data already calculated from the June crawl.

In order to determine how many samples are needed to reach a 5% error margin, we used the already calculated barrier probabilities. If the complete set of samples were not enough to reach an error margin of 5%, the collected samples were copied⁴. In this way we could simulate more samples than we actually have. By doing this, we can get an estimate of how many samples that are needed to reach the predefined error margin. An outline of this simulation can be seen in figure 3 where the x-axis outlines how many samples were needed while the y-axis outlines the number of sites requiring this number of samples.

An error margin of 5% (0.05) within 95% confidence interval might in some cases be sufficient, because the resolution of the UWEM scorecards, presented in table 1, is 0.25.

⁴ We assumed that the sample distribution would not change significantly from the 100 samples we already had, when copying.

However, if the average barrier probability for a site fell between two scores, it would be random

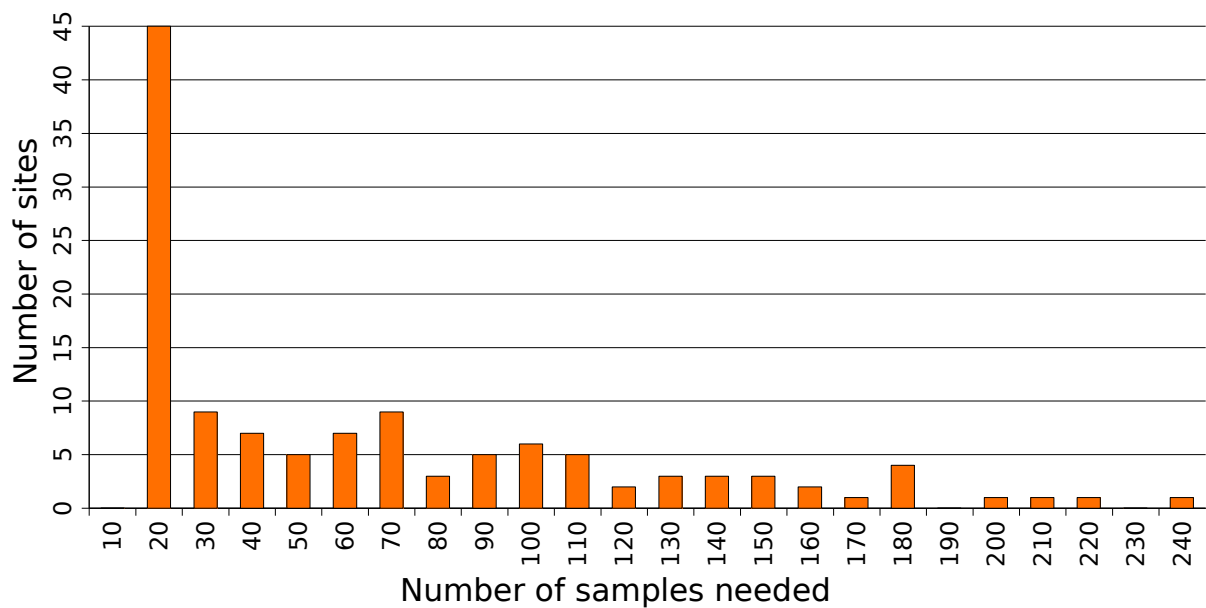


Figure 3: Adaptive sampling

which of the scores that was chosen, so in this case we would need other measures to improve the reliability of the results, like e.g. a sliding average over a time period to avoid random fluctuations between monthly test runs.

We can see from the simulations that on average only 61 samples is needed to achieve the required precision, in contrast to the 100 samples used so far. The highest number of samples to achieve this precision was 211 samples. This means that adaptive sampling to a 5% error margin⁵ would give 39% speed improvement compared to what is currently implemented, and it would be easier to disclose an overall sampling error margin, than one per web site. We required at least 10 samples to accept the result, which is why the histogram for 10 samples is empty, whilst there is a peak at 20 samples.

One site did not achieve the required error margin, however it had only two URLs defined, which can not be regarded as a reliable result. Most probably we had a problem crawling this site due to JavaScript or Flash. We plan to investigate all web sites with few pages identified (e.g. less than 10) and publish the results if the investigations shows that the data is reliable.

This simulation clearly shows the benefit of adaptive sampling, since the error margin varies between 0 and 7% without adaptive sampling, which means that the worst case error margin can be improved by 2% at the same time as the average number of samples is decreased by 39%.

45 web sites (37% of total) required less than or equal to 20 samples to achieve a 5% error margin, which means that it should be feasible to perform a combination of manual and automatic assessments to a useful error margin for many web sites.

⁵ Note that 95% CI means that 5% of the data on average still will be outside the error margin of ± 0.05 .

3.4 Limitations of automatic accessibility monitoring

The large scale accessibility assessment performed by EIAO relies only on automatic tests. A manual approach would not be feasible for this number and frequency of evaluations. Automatic evaluation can never reach the same quality level as an expert evaluation regarding the completeness and validity of results because many accessibility issues need human judgment (e.g. the appropriateness of alternative texts).

Nevertheless, automatic monitoring can provide useful results. The barriers that are detected automatically serve as a kind of *indicator for the accessibility* of the web site. The rationale behind this is that web developers who make their sites accessible for people with disabilities are likely to take into account more aspects of accessibility, not only the ones that can be checked automatically. On the other hand if there are issues reported by the automatic evaluation there might be more barriers.

An important feature of the results presented by EIAO is their *repeatability* and *comparability*. As all evaluations are carried out with the same setup the results for several web sites or for different versions of the same web site can be easily compared. Whereas for expert evaluations the repeatability is lower because "subjective" judgments are involved.

The probabilistic model still has some limitations because the parameter estimates are not yet refined. EIAO is conducting user testing surveys to improve the parameters and thus the usefulness of the aggregation model. The first promising results are reported in [15].

4 CONCLUSIONS

We have demonstrated some early results from the EIAO Observatory and have shown that a random walk approach for random selection of web pages, together with an adaptive sampling strategy to a given error margin, can provide repeatable and coherent automatic accessibility assessment results to a required error margin with the UWEM aggregation model. Simulations have shown that we can achieve an error margin of 5%, which is 2% better than the current worst case error margin with 39% less samples than we have today. Another nice conclusion is that for many web sites the required number of web pages to sample is less than 20 pages, which means that it should also be feasible to perform expert evaluations using a combination of automatic and manual assessments using the UWEM aggregation method and still get results that are within a decent error margin given 95% confidence interval for many web sites. In addition, it is possible to compare results between different monthly tests, to perform an analysis of the reason for change in score values, i.e. whether the reason is a real change in accessibility, "strange" sample sets that are outside the confidence interval or any other reason.

We are continuously working on improving the Observatory. The main efforts are currently on improving scalability, test precision, adaptive sampling, scoping rules and covering the entire automatable test set in UWEM 1.0. In addition we are working on improving the UWEM methodology.

REFERENCES

1: WAB cluster (EIAO, BenToWeb and SupportEAM projects) Unified Web Evaluation Methodology (UWEM 1.0) 2006 URL: http://www.wabcluster.org/uwem1/UWEM_1_0.pdf

- 2: E. Velleman et al. WAB cluster 2006 URL: <http://www.wabcluster.org/>
- 3: EIAO Consortium European Internet Accessibility Observatory 2006 URL: <http://www.eiao.net>
- 4: W3C Web Content Accessibility Guidelines URL: <http://www.w3.org/TR/WAI-WEBCONTENT/>
- 5: World Wide Web Consortium URL: <http://www.w3.org/>
- 6: Web Accessibility Initiative - WAI URL: <http://www.w3.org/WAI/>
- 7: Mikael Snaprud, Nils Ulltveit-Moe, Anand Pillai and Morten Goodwin Olsen , A proposed architecture for large scale web assessment, ICCHP 2006 URL: <http://eiao.net/publications/>
- 8: A.B. Pillai HarvestMan 2006 URL: <http://harvestman.freezope.org>
- 9: P. Nalevka The Relaxed HTML validator 2006 URL: <http://badame.vse.cz/validator>
- 10: Schematron URL: <http://www.schematron.com/>
- 11: Monika Henzinger, Link Analysis in Web Information Retrieval, 2000 URL: <http://www.research.microsoft.com/research/db/debull/A00sept/henzinge.ps>
- 12: Christian Thomsen, Torben Bach Pedersen, Building a Web Warehouse for Accessibility Data, 2006, to appear in DOLAP'06
- 13: Nomenclature of territorial units for statistics - NUTS URL: http://ec.europa.eu/comm/eurostat/ramon/nuts/home_regions_en.html
- 14: Evaluation and Report Language URL: <http://www.w3.org/TR/EARL10-Schema/>
- 15: Cristian Bühler, Helmut Heck, Olaf Perlick, Annika Nietzia and Nils Ulltveit-Moe, Interpreting Results from Large Scale Automatic Evaluation of Web Accessibility, ICCHP 2006 URL: <http://eiao.net/publications>