

Project ref. no.	EDC-52005 ELECT/27644
Project title	ELECT – The European Localisation Exchange Centre

Deliverable status	Restricted
Contractual date of delivery	Month 24, January 2004
Actual date of delivery	Month 24, January 2004
Deliverable number	D1.1-2
Deliverable title	<i>ELECT Expert Workshop Proceedings Development of global web sites – Internationalisation</i>
Type	Report
Status & version	Final
Number of pages	
WP contributing to the deliverable	WP1
WP / Task responsible	LRC
Author(s)	Reinhard Schäler
EC Project Officer	Erwin Valentini
Keywords	ELECT, expert, workshop, proceedings, internationalisation
Abstract (for dissemination)	Proceedings of the ELECT Expert Workshop, September '03



Expert Workshop

Development of global web sites
Internationalisation
Current and future issues

06 September 2003
Atlanta (Georgia, USA)

This report was compiled by Reinhard Schäler following the expert workshop organised by the LRC as part of the European ELECT project. While the author gratefully acknowledges the input of the participants of the workshop and expresses his thanks to them for their contribution, the content of this report is solely the responsibility of the author.

ELECT is funded by the European Union's eContent programme.



© 2004 Localisation Research Centre (LRC), Department of Computer Science and Information Systems (CSIS), University of Limerick (UL), Limerick, Ireland.

Executive Summary

The ELECT expert workshop on the international development of digital content took place in connection with the 24th Internationalization and Unicode Conference on 06 September 2003 in the Double Tree Hotel, Atlanta, Georgia, USA. The participants were Richard Ishida, Martin Dürst and Reinhard Schäler.

Richard Ishida is based in the United Kingdom. He works for the World Wide Web Consortium and is the team contact for the Internationalization Working Group. He is chair and team contact for the GEO task force (Guidelines, Education and Outreach). He is also the co-chair of the Internationalization & Unicode Conference, and on the board of the International Conference on Usability and Internationalization.

Martin Dürst joined the W3C Team at Keio University (Japan), SFC, in December 1997 to work on Internationalization. He is now a Visiting Scientist at the Massachusetts Institute of Technology (MIT) Laboratory for Computer Science (LCS). Prior to joining W3C, he was at the University of Zurich, Department of Computer Science, and has been an active participant within the HTML and CSS Working Groups as an invited expert on internationalization.

Reinhard Schäler is the director of the Localisation Research Centre (LRC) at the Department of Computer Science and Information Systems. He is the co-ordinator of the ELECT project.

A number of issues were highlighted and solutions discussed during this workshop, amongst them:

- The representation of text encoding and the general move towards Unicode
- The problem of the so-called *tag soup* and the move towards style sheets
- The support of non-western typography in CSS
- The support of right-to-left scripts in an HTML environment

Table of contents

EXECUTIVE SUMMARY	4
INTRODUCTION	6
DEVELOPMENT OF GLOBAL WEB SITES: ISSUES	6
REPRESENTATION OF TEXT ENCODING	6
<i>Example</i>	6
REPRESENTATION OF DATA	7
<i>Example</i>	7
<i>Recommendation</i>	7
LANGUAGE AND LOCALE NEGOTIATION	7
<i>Example</i>	7
<i>Suggestion</i>	7
ACCESSIBILITY	7
<i>Recommendation</i>	8
HUMAN COMPUTER INTERFACE (LEGAL AND FINANCIAL RESTRICTIONS)	8
<i>Recommendation</i>	8
TAG SOUP AND BLOAT	8
<i>Example</i>	8
UNSTRUCTURED HTML	8
<i>Effect</i>	8
CASCADING STYLE SHEETS (CSS)	8
<i>Solution</i>	9
BI-DIRECTIONAL EDITING IN HTML	9
<i>Solution</i>	9
INTEGRATION OF MULTIMEDIA	9
<i>Outlook</i>	9
SUPPORT INFRASTRUCTURE	9
GUIDELINES	9
<i>Recommendation</i>	9
BUYER/SELLER AWARENESS	9
<i>Recommendation</i>	10
OUTLOOK	10
REFERENCES	12

Introduction

Most of the basic issues in relation to the internationalisation of web content have been solved or are, according to the experts, close to a solution – at least in principle. These solutions are described in documents published by the World Wide Web Consortium (W3C) on their Architecture Domain site <http://www.w3.org/International/>. This site contains specifications, articles, notes, tutorials, tests, tools, and introductory material as well as a section on frequently asked questions (FAQs).

The most recent overview of web internationalisation is that provided by Tex Texin and Yves Savourel. As part of the EU-funded ELECT project, the LRC also developed course material on the development of global web sites.

While in some cases solutions are proposed in this report, in others the report only provides examples or recommendations.

Most issues identified in this report address one of the following areas:

- Digital content development
- Representation of data
- Structure of data (text vs format)
- Inclusion of graphics
- Scripting (e.g. CGI)

However, as some issues are not confined to one particular area we have decided to list issues without grouping them under any one particular header. The only exceptions are those listed under the header *Support Infrastructure*.

Two of the most basic, yet largely unresolved issues in multilingual digital content production and processing are:

- A standard approach to the mark-up and identification of localisable data
- A standard approach to the automation of the localisation process

Both of these issues are currently being addressed by two technical committees (TC) within the OASIS consortium, namely the TC on the Localisation Interchange File Format (XLIFF) and the TC on Translation Web Services (Trans-WS). They are not being dealt with in this report.

Development of global web sites: issues

Representation of text encoding

There is an industry-wide movement towards the use of Unicode¹. Although some content developers and tools providers do not support it yet, there is evidence to suggest that this is happening.

Example

TRADOS tools did not support Unicode fully for quite some time.

¹ Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language. For more details on Unicode: www.unicode.org

Representation of data

There are cases where it is necessary to store data using one encoding scheme or one set of conventions and to display it using another encoding scheme or other conventions. It can even be necessary for the same data to be processed in different ways because of different conventions and legal requirements in specific *locales*².

Example

In different *locales*, basic data such as dates, time and numbers are displayed in different ways. In addition, applications might have to function in different ways and in accordance with the conventions and regulations of a specific *locale*. This can include address checks, payment procedures and legal regulations.

Recommendation

Self-assessment checks could be developed to enable digital content developers and localisation service providers to establish whether and to which extent their domain-specific content is ready for deployment in specific locales.

Language and locale negotiation

When requests to serve digital content are submitted to web servers (and similar systems) it is often difficult to establish what the preferred locale and language settings of a particular user are. Just because a user is located in a specific country or because he is using a computer running an operating system in a specific language does not necessarily mean that he would prefer to have web pages served to him in the language and with the settings of that specific locale.

Example

Business travellers and tourists often have to use systems using the language and locale settings of the host country rather than their preferred language and locale. Although web sites often allow users to switch between different languages, changing these settings generally does not affect the *locale* settings, which are, in most cases, handled by the operating systems running on the local computer.

Suggestion

Language and *locale* negotiation should be determined by preferences that a user can select at any given time. They should not be pre-determined by the local settings of the computer or by the system serving digital content. While the option of changing the language of the content served is a great advantage over monolingual content, a change in language should also affect the way the information is formatted and presented according to specific *locales*.

Accessibility

Modes of input (keyboard, mouse, speech etc.) and output (printer, paper sizes, speech etc.) are very important to ensure accessibility of digital information for people with different abilities and for people using different types of devices, from mobile phones to Personal Digital Assistants (PDAs), other handheld computing devices and even refrigerator doors.

Current efforts to ensure accessibility of digital content concentrate on language and locale settings on specific platforms and for specific devices.

² A *locale* is a generic term indicating a set of attributes related to language and other regional/ethnic preferences. Examples include currency symbols, date and time formats, calendar types, number formats, default character encoding, and keyboard layouts.

Recommendation

As the variety of computing devices grows, digital content should be made platform independent. It should also be made accessible to people using input and output devices according to their own personal preferences and abilities. The use of mark-up languages such as XML can make the development and the deployment of multimodal multilingual digital content significantly easier.

Human Computer Interface (legal and financial restrictions)

Some human computer interface (HCI) issues are still difficult to address given the current legal and financial restrictions on the worldwide trading of goods. It is not uncommon to find that users can only access digital content or participate in digital trading if they are located in a specific geography or legislature or have a bank account in a given country.

Often, user interfaces do not take account of these restrictions and so users are not aware from the outset that they might not be permitted to take advantage of particular offerings.

Recommendation

Terms of use, trading policies and restrictions should be stated clearly from the outset to any users trying to access digital content or to trade goods electronically.

Tag soup and bloat

There is an industry-wide move towards the use of style sheets (CSS) and away from the rather messy HTML-only format. There are significant advantages in using highly structured web pages and sites as changes in the original site, which should be replicated in the language versions of the same site, can be noted much easier. The more structured the source content, the easier it is to localise.

Some older versions of browsers might not be able to show sites using style sheets. In this case, a message should be returned to those users asking them to update their browser.

Example

ESPN, Microsoft's sports Internet channel, becoming aware of the advantages of structured coding of web pages, changed the whole site and implemented style sheets.

Unstructured HTML

Some content developers are moving from the rather unstructured HyperText Markup Language (HTML) towards the Extensible HyperText Markup Language (XHTML), which is XML-based (Extended Markup Language) but reproduces, subsets and extends HTML 4. A further move from XHTML towards pure XML in the near future is anticipated and, with it, a growing use of the Extensible Style Sheet Language Family (XSL), which are the recommendations for defining XML document transformation and presentation. XSL Transformations (XSLT), the language for transforming XML documents, will play an increasingly important role.

Effect

As the move from HTML to XHTML and XML progresses, development and localisation tools and technologies will have to be adapted to the new formats.

Cascading Style Sheets (CSS)

The widely used Cascading Style Sheets (level 1) lack the ability to support non-western typography, including vertical text, kashida³-based justification, controllable line-breaking, Ruby⁴ characters and others.

³ Arabic elongation character

Solution

The migration from CCS1 to CSS2 or, even better, CSS3 should solve most of the problems experienced in relation to non-western typographies.

Bi-directional editing in HTML

HTML is a Left-to-Right environment, i.e. it was developed with the traditional western and latin-based writing system in mind. It only caters for languages using a Right-to-Left (RTL) writing system, such as Arabic and Hebrew, in limited ways through the use of attribute names and attribute values. There is no bi-directional algorithm available to distinguish between mark-up of text and added functionality.

Solution

Better support for bi-directional editing is needed. This could be achieved with a customised Emacs tool.

Integration of multimedia

The integration and, where necessary, adaptation of images in digital content and particularly in web sites is generally problematic.

Outlook

Scalable Vector Graphics, or SVG, is a highly sophisticated language for describing two-dimensional graphics and graphical applications in XML and is particularly well suited for embedding graphics in multilingual digital content.

Support infrastructure

Guidelines

There is a vast amount of information available on the development of multilingual digital content in a web environment and the localisation of such material. However, this material is generally not well structured or addresses the requirements of specific development and implementation environments. For example, guidelines are available for the Apache server but these cannot be applied across the board. There is a need for the production of well structured, authoritative, agreed and widely recognised guidelines. These should include:

- Language negotiation
- Folder / directory structure
- Language selection

Recommendation

Guidelines for the development of multilingual digital content and the localisation of digital content should be developed by either European consortia or independent, well respected organisations in cooperation with the stakeholders.

Buyer/Seller awareness

There is a general lack of awareness in relation to multilingual and multicultural issues among the buyers and sellers of digital content applications, i.e. operating systems, word processors, browsers, digital content management systems etc.

When pointing to pages in different languages links will be displayed in the font of the target languages. This causes, potentially, two problems:

⁴ Ruby are small characters used for annotations of a text. They are positioned at the right side for vertical text and on the top for horizontal text. They indicate the reading (pronunciation) of ideographic characters and are used in Japan and in China in many publications, such a books and magazines.

- If the target font is not coded in Unicode, the font cannot be displayed correctly;
- If the target font is coded in Unicode the corresponding fonts have to be installed in order to display the text correctly. (A full set of Unicode fonts is not installed “out of the box”, automatically, when operating systems and applications are installed on hardware.)

Buyers and sellers are often not aware of these potential problems. Even if they are aware of them they do not always have access to the necessary expertise to address them appropriately.

It becomes more urgent to address this problem every day as more people make use of global networks.

Recommendation

There are a number of possible actions to address this problem:

- A basic set of Unicode compliant fonts covering the full range of code points could be installed automatically on all hardware.
- A basic certification system could be implemented giving potential buyers and sellers an indication to which extent operating systems and applications are multilingual ready.
- The Internationalized Resource Identifiers (IRI) proposed by Dürst and others are also addressing this problem

Outlook

Adaptation of digital content has to happen at different levels and beyond that of the adaptation of language and *locale*. Social, technical and financial issues have to be taken into account when making digital content available simultaneously to users independently of their geographical location and of their cultural and linguistic background.

Independent, respected and widely supported bodies such as the World Wide Web Consortium and OASIS have established guidelines for the development of multimodal, multilanguage and multicultural digital content. These guidelines need to be further developed.

Self-assessment checks or checks by independent bodies should be developed and their results published with the digital content so that users can become immediately aware of the grade of adaptation of particular content.

Most of the issues in relation to the publication of digital content in a global environment have been resolved technically. However, what is technically possible has not yet filtered through to the majority of implementations.

While this report of the ELECT expert workshop on the development of global web sites describes many of the issues, recommendations and solutions to the publication of digital content in a global environment, more detailed and ongoing studies are required to reflect the current state of implementation of new approaches and technologies.

These studies should address, among others, the following questions:

- Which needs have been recognised but not yet been addressed?
- Which prototype solutions have been attempted?
- Which technologies are available but not yet deployed universally?
- What is current good practice in global digital content publishing?

Regular and established conferences such as the annual event organised by the LRC and the bi-annual International Internationalisation and Unicode Conference should be used to disseminate the result of these studies.

They should be followed up by medium to long-term educational and research programmes at universities and more short-term training initiatives for industry.

References

Bos, B., *Cascading Style Sheets*, Home Page. <http://www.w3.org/Style/CSS> (Last consulted: January 2004)

Dürst, M., *The next topics for www internationalization*, Position paper for the Workshop on Internationalization at the 5th WWW Conference in Paris, 06 May 1996. <http://www.w3.org/International/martin.duerst.html> (Last consulted: January 2004)

Dürst, M., *Ruby in HTML*, <http://www.ifi.unizh.ch/groups/mml/people/mduerst/ruby.html>, last updated 14 May 1996. (Last consulted: January 2004)

Dürst, M., *Internationalized Resource Identifiers: From Specification to Testing*, Proceedings 19th International Unicode Conference, San Jose, September 2001, <http://www.w3.org/2001/Talks/0912-IUC-IRI/paper.html> (Last consulted: February 2004)

Dürst, M., and M. Suignard, *Internationalized Resource Identifiers (IRIs)*, draft-duerst-iri-06, February 15, 2004, <http://www.rfc-editor.org/internet-drafts/draft-duerst-iri-06.txt> (Last consulted March 2004) (Expires: August 15, 2004)

ELECT, *Foundation course in web development for the global market*, Localisation Research Centre, 2003.

Harvey, G., GNU Emacs, Free Software Foundation (2004), Inc., 59 Temple Place - Suite 330, Boston, MA 02111, USA. Available at: <http://www.gnu.org/software/emacs/emacs.html> (Last consulted: February 2004)

Nelson, P., *Justifying Text using Cascading Style Sheets (CSS) in Internet Explorer 5.5*, Microsoft Corporation. Available at: <http://www.microsoft.com/middleeast/msdn/JustifyingText-CSS.aspx> (Last consulted: January 2004)

SVG Working Group, *Scalable Vector Graphics (SVG) 1.1 Specification*, W3C Recommendation (14 January 2003). Available at: <http://www.w3.org/TR/SVG/> (Last consulted: January 2004)

Texin, Tex, and Yves Savourel, *Web Internationalization – Standards and Practices*, 24th Internationalization and Unicode Conference, Atlanta, GA, September 2003. Available at: <http://www.xencraft.com/resources/webi1&ntutorial.pdf> (Last consulted: January 2004)