

Online version available from http://www.w3.org/2006/Talks/fundamentos-web-ri/





## Localization

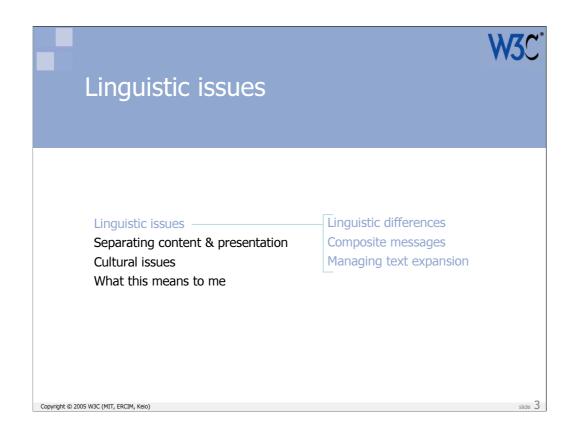
The **adaptation** of a product, application or document content to meet the language, cultural and other requirements of a specific target market.

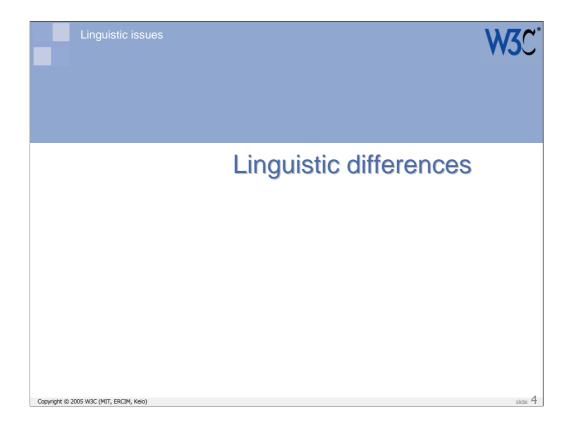
## Internationalization

The **design and development** of a product, application or document content that **enables** easy localization for target audiences that vary in culture, region, or language.

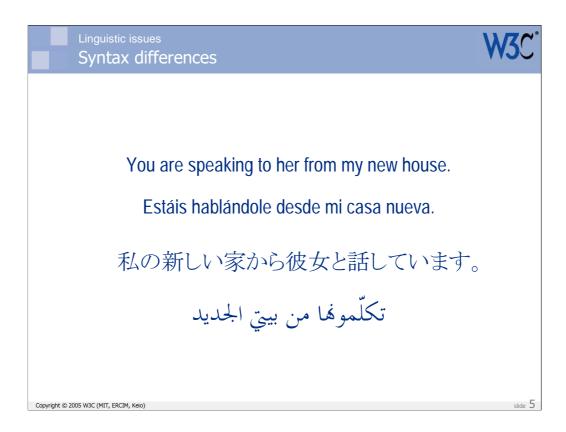
Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 2

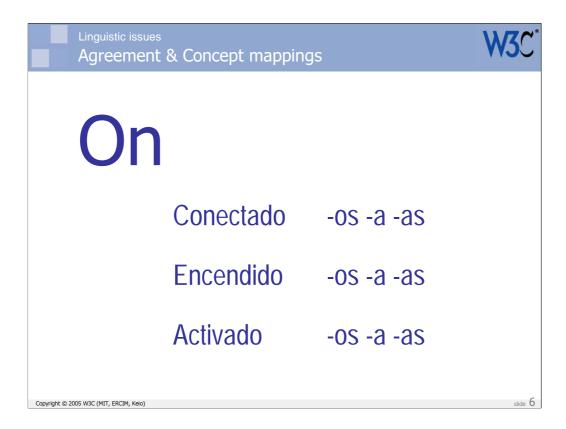




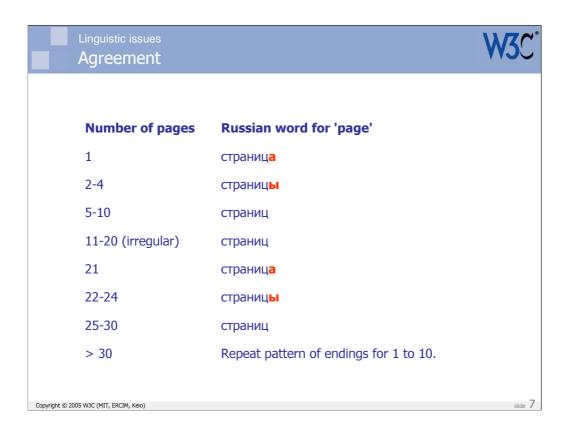
In this first section we will look at a few ways in which languages differ, and then see those differences causing practical issues for localization where the developer/designer has not thought about internationalization.



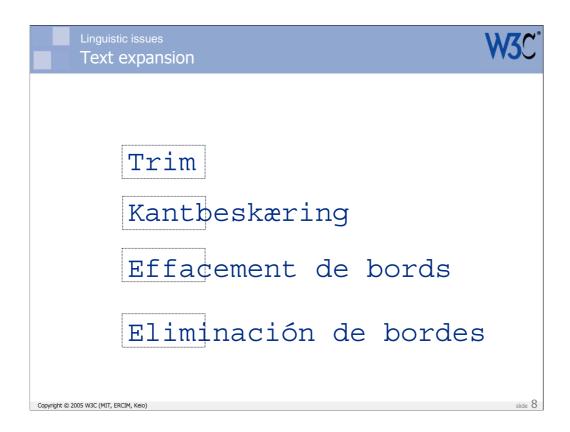
This shows four different ways of writing one idea. In each case the order of 'words' and the number of 'words' is different.



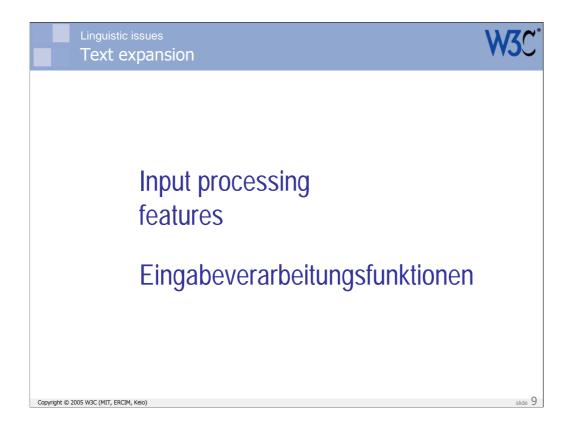
This slide shows how the English word 'On' can map to three different words in Spanish. And then there are the masculine, feminine and plural forms of agreement that change the shape of the word according to its context.



In Russian there is a complex plural system. Apart from the irregular teens, the word endings are applied in a rotating way.

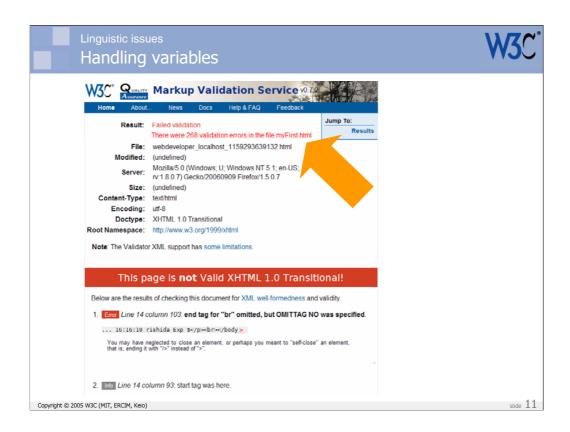


This slide introduces the idea that terms or labels can be of widely differing lengths in different languages.

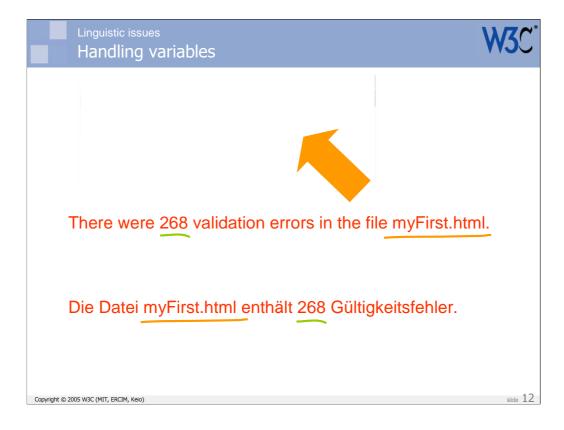


In languages such as German, Dutch or Swedish it is common to find English 'compound nouns' expressed as a single, long word.

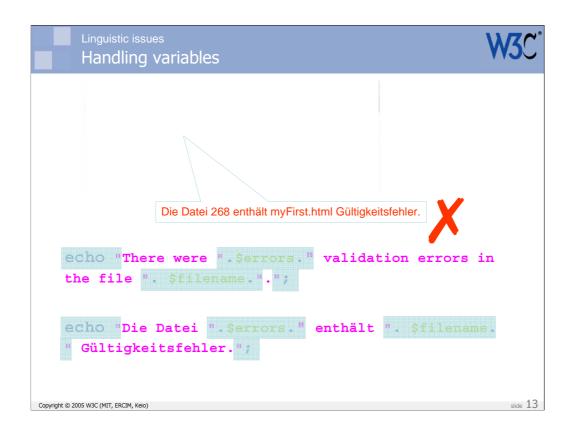




For this slide we imagine that the W3C Validator is altered slightly so that it tells you how many validation errors are in your file. It will do this using a 'composite message' whose parts are assembled using PHP code as the page is served.

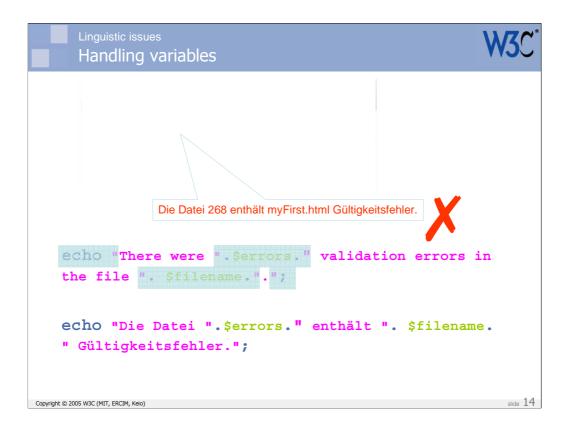


In the German translation, the order of the two variables may need to be changed.

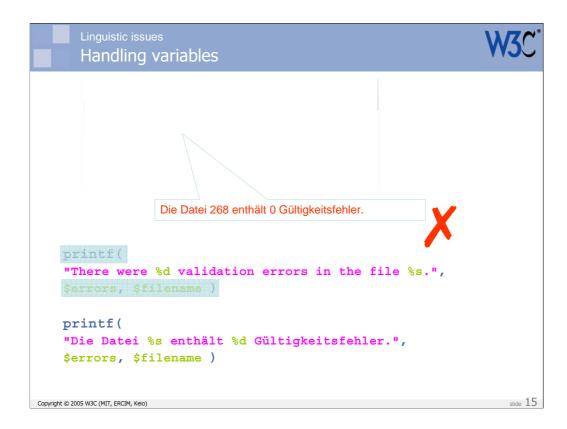


Typically translators have no access to the actual code, to avoid them introducing bugs into the page. Either the text is extracted or a translation tool masks the code.

Although we are fortunate that we were able to add words after the second variable, due to the English string containing a period, this still didn't produce the right result. The German reads "File 268 contains myFirst.html validation errors."

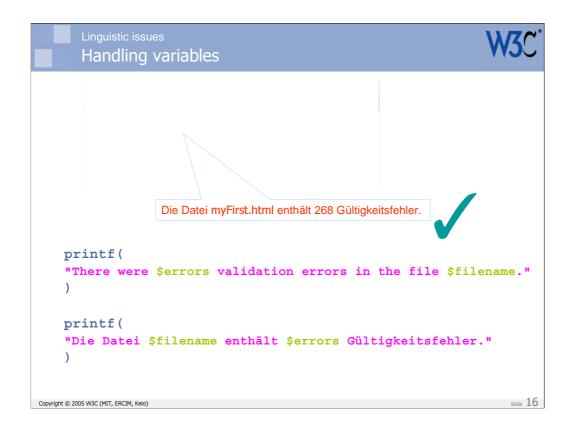


The reason is that the translation process didn't switch the order of the variables.



So next we try using a printf statement. This has the benefit that text and variable locators all sit within a single string, and the translator can access the items they want to reorder.

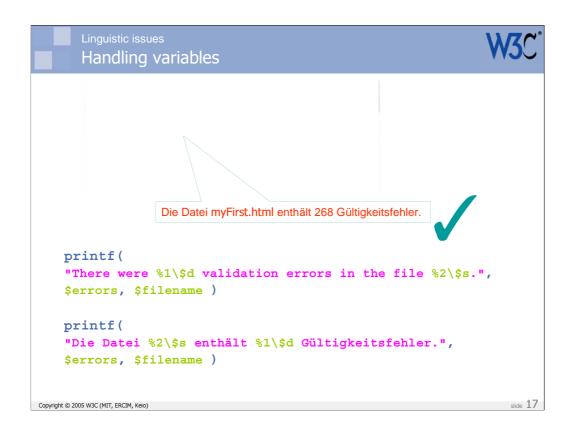
Unfortunately, this doesn't help, since PHP still replaces the variables in the string in the order of the variables cited in the following parameters to printf. This causes the 268 to be shown instead of the filename, by converting the integer value to a string. It is unable to find an integer value in the file name, and so presents us with the zero for the number of errors.



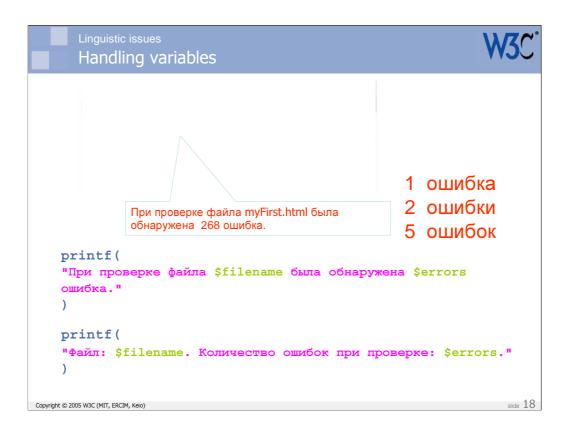
By embedding the variable names directly in the printf string, as shown in this slide, we finally achieve the desired result in German.

Nota bene: Successful, or at the very least, cost effective localization in this case is down to the **designer/developer** understanding the potential pitfalls of various approaches to coding. It is not the job of the localization vendor to get this right. It needs to be done as the initial content is created!

You should also be very careful of the assumption that 'This doesn't affect me, since we don't translate the content I develop.' I have seen many, many cases where the thing being developed was later so successful that people wanted to take it to other regions, only to find that they ran into major difficulties because of issues with the translatability of the code or content. It's best to just do it right from the start.

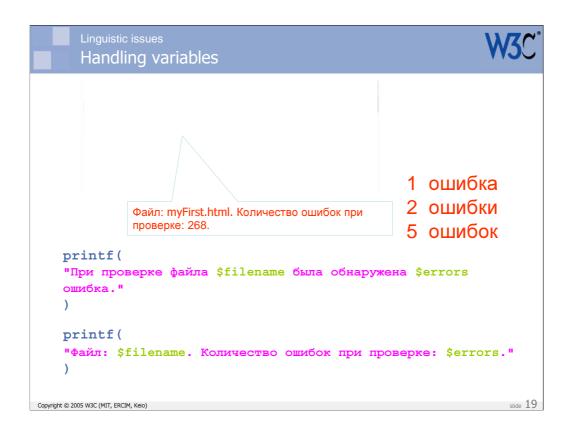


By the way, there is a way to produce the right effect while using the %d and %s variable markers in a PHP string, but it involves a slightly more complex syntax. This is shown in the above slide. The numeric markers refer to the relevant variable in the parameters that follow the string, even after reordering.

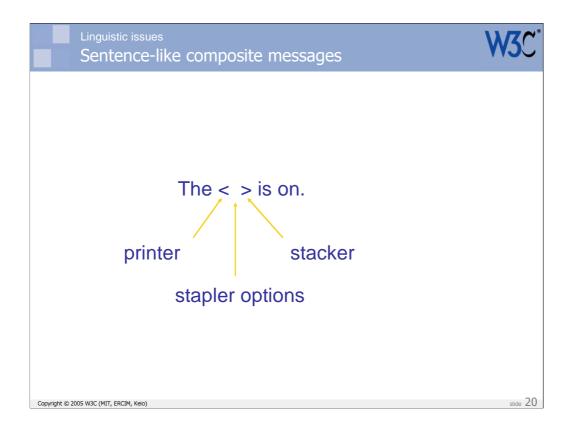


So now we know how to code this type of text in PHP... or do we?

Let's think back to our example of how plurality works in Russian, and we realize that we still have a problem for that language. We only have a single string and it can only be translated one way – yet the Russian requires three variants of the word ошибка, depending on the number that precedes it.

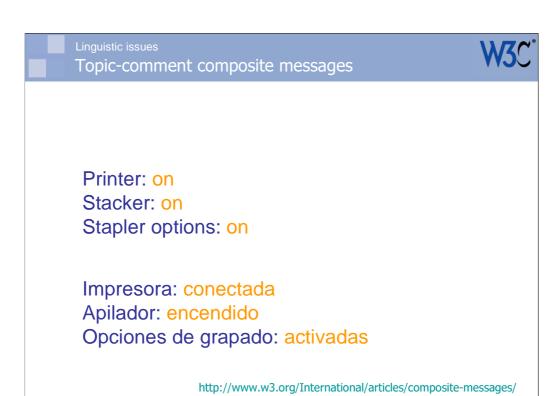


To deal with this, the Russian translator would probably resort to a completely different structure for the text, essentially equivalent to "File: X. Validation errors: Y". This approach requires only one form of ошибка in the invariable string. This is an example of what I call a 'topic-comment' composite message.



So we are beginning to see here that there are two distinct types of composite message. The first is based on a sentence-like approach, and the invariant string can be difficult to translate in some circumstances because of the need for agreement or different word mappings.

In the example above, 'The' should be translated 'el', 'la', or 'las' in Spanish, depending on what word follows it. Also the word 'on' should be translated using three different Spanish words (with different endings).



The other approach to designing composite messages is what I like to call the 'topic-comment' approach: you state a topic, then you say something about it.

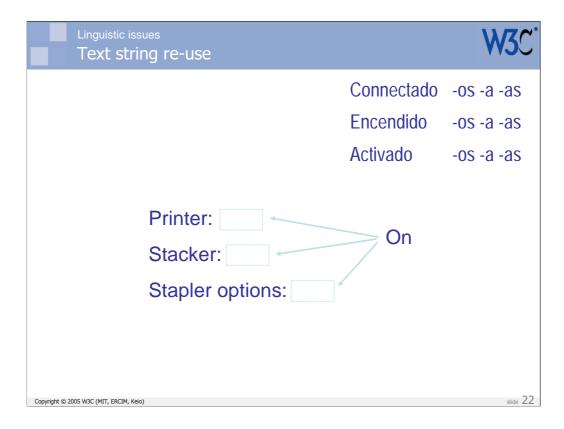
Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 21

This approach works much better for the previous slide, since you each comment you associate with a topic can use a different word with the appropriate word endings.

There is a little more to this theory of composite messages than we have mentioned so far, but you can get more information from the W3C Internationalization site at the following URI:

http://www.w3.org/International/articles/composite-messages/ .



I should, however, mention just one other point. Many designers/developers looking at the English topic-comment arrangement on the previous slide might think to themselves that they could save a little bandwidth by reducing all those instances of the word 'On' to a single string that is used for all comments, ie. they want to re-use strings.

```
cy>Printer: <?php print getStatus($printer);?>
stacker: <?php print getStatus($stacker);?>
stacker: <?php print getStatus($stacker);?>
stapler options: <?php print getStatus($soptions);?>

<?php
function getStatus (device) {
    if ($device.status!= 0) { return 'On'; }
    else { return 'Off'; }
    }
}
</pre>

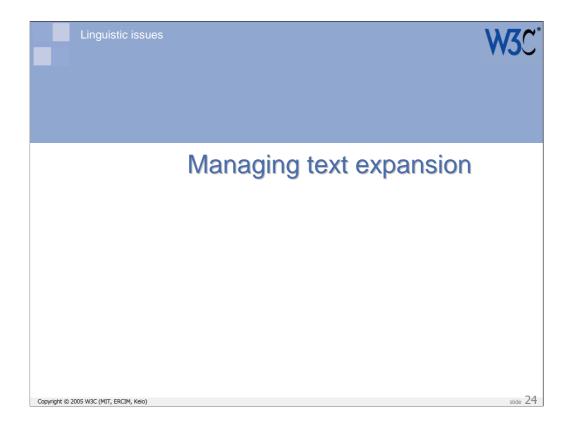
cyproprio ($ 2005 W3C (MT, ERCIM, KeO)

bttp://www.w3.org/International/articles/text-reuse/
```

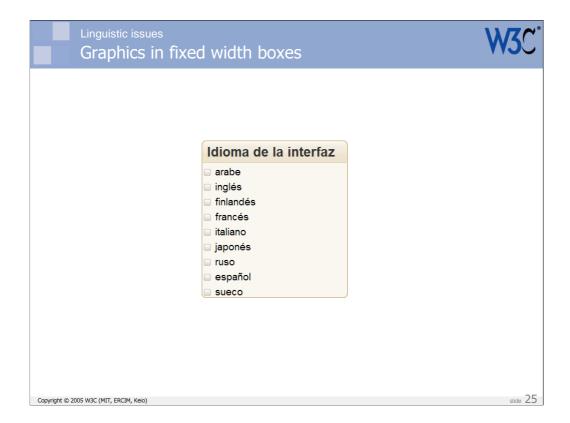
Tempting as this idea may appear, it will unfortunately introduce insurmountable problems for translation, since the comment is likely to require different agreement forms at the least, and possibly different words altogether, depending on the context.

This slide shows an example of how such a problem may come about by returning the same text from a function for each comment.

Note that I do not want to rule out string re-use altogether – there are situations where it is a sensible approach. But re-use must not occur across different contexts. For more information about this, see the W3C Internationalization article at  $\frac{\text{http://www.w3.org/International/articles/text-reuse/}}{\text{text-reuse/}}$ 



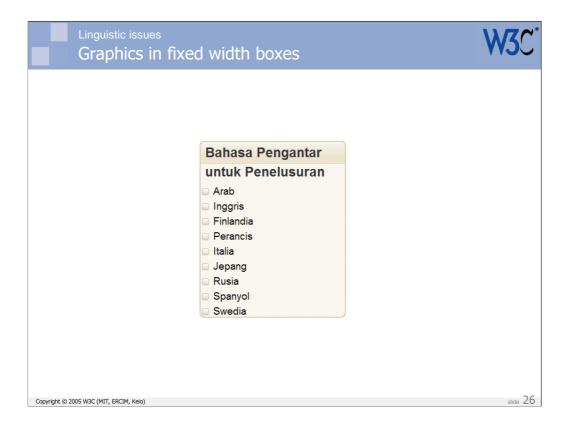
Now we switch to a very different topic area, that has more to do with the visual layout of the page than the composition of the text.



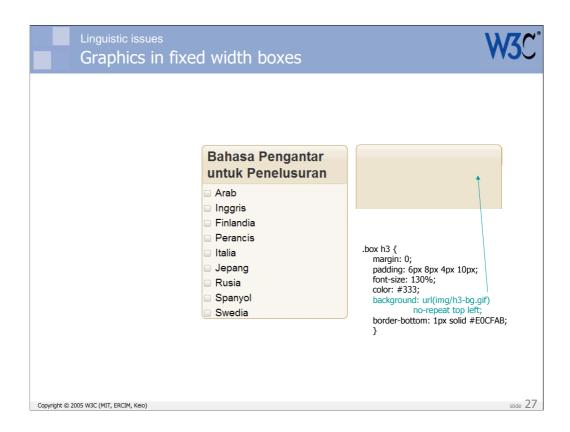
Let's assume that we want to implement a **fixed-width** box on our page. The text can expand downwards, but not sideways.

Let's also assume that we want a background with a nice gradient behind the title of the box, and that the background has a line across the bottom.

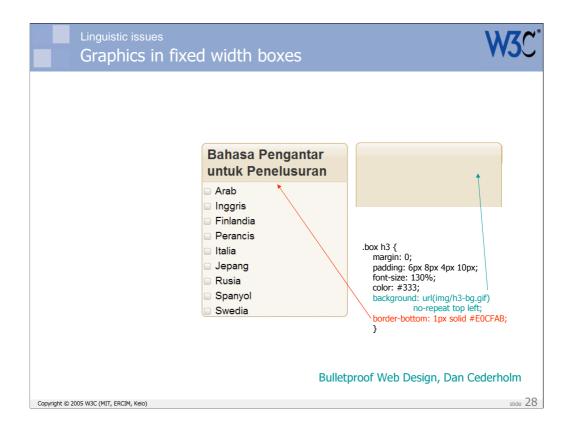
(This slide in Spanish has the title 'Interface Language', and a list of radio buttons to select a language.)



As our text expands during translation into Malay, the title occupies two lines. Unfortunately the graphic used for the gradient background is only one line deep, and things now begin to look a mess.



A way to approach this issue is to use a graphic that is three or four lines deep behind the title. By attaching the graphic using the CSS background property, only the amount needed to view the title will actually be shown.

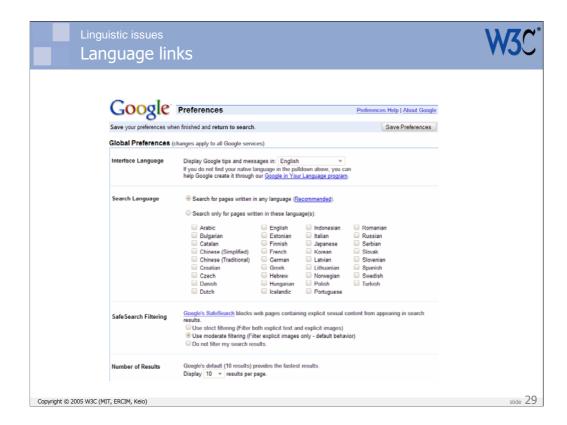


To get the line to appear in the right place, we simply create it as the bottom border of the heading.

This example uses a technique (and the exact same code and graphic) described in Dan Cederholm's book, Bulletproof Web Design (although the text is borrowed from Google's language preferences). This is significant! Dan is not writing about internationalization per se – he is more concerned with people pumping up the text size for accessibility reasons. It just so happens, however, that the same approach helps with localizability.

This is an example that you don't necessarily have to learn new information to deal with internationalization issues – just following existing best practices can be the key in many cases. Note again, however, that we are still talking about the **design and development** of content – not about work that the localizers will do!

Dan's book contains several other recommendations that will benefit internationalization.



Note, in passing, an issue related to the Google text I used in the previous example. The dialogue allowed you to select a different language for the user interface from a pull-down list, presumably assuming that your reason for changing was that you couldn't read the current language.

The issue for me is that the names of all the languages are in the language of the current page. Let's assume, for example, that a curious person wanted to see what the interface looked like in Persian, so they selected that language from the list and clicked on the 'Save Preferences' button.



Assuming that they would be able to find their way back to the appropriate dialogue box to get back to English (which would require them to remember which link to hit on the thankfully uncluttered Persian Google home page), that they can remember which is the required select list, and that they can do so in spite of the mirror-imaging of the page when using Arabic script, they would then be faced with what you see on the next slide.



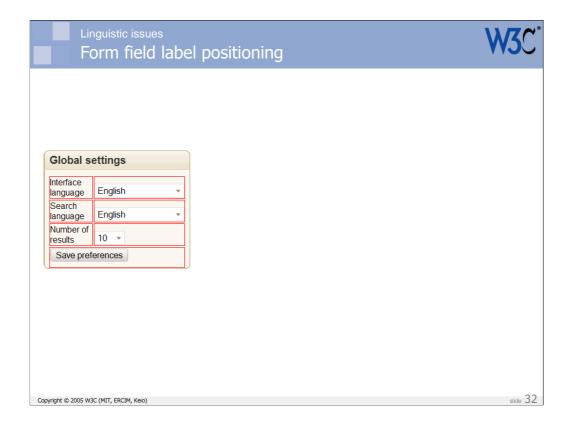
Note that the names of languages are all in Persian, and are sorted by Persian rules.

Which selection would get you back to English ??

(Hint: if you want to explore like this, use a different tab or window for your explorations, and leave the original dialogue available in another for when you want to reset to your current language.)

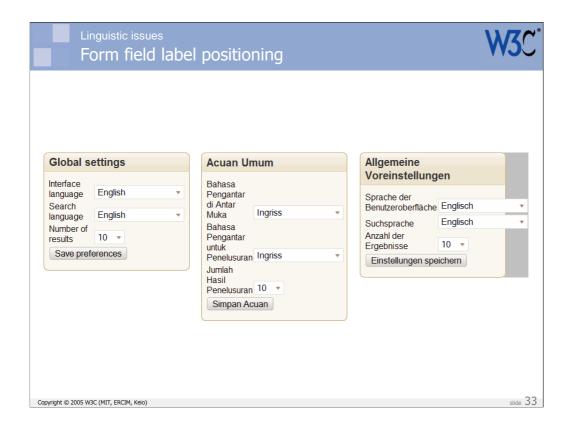
Of course, the point is really that a Persian person taken to the English site may have as much trouble finding their way to the appropriate user interface language as the curious explorer does in getting back. In my opinion it would help a great deal to write each language name in its own script and language. You can read more about that in the W3C Internationalization article at

 $http://www.w3.org/International/questions/qa-navigation-select \; . \\$ 



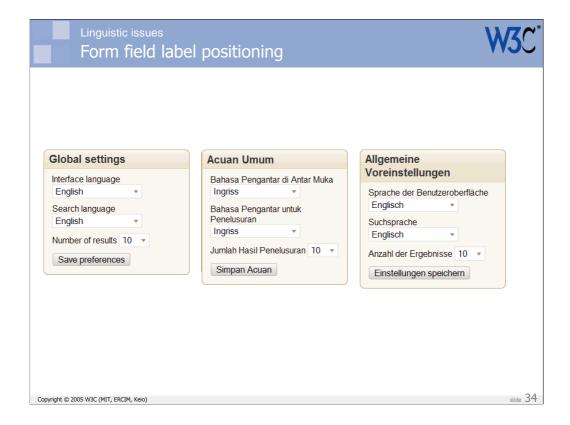
Lets take a moment to explore another potential issue related to the length of text in translation.

Let's continue to assume a situation where text appears in a fixed width box. We will apply the same approach we discussed earlier to deal with the title of the box. The issue this time will be that we have used a table to apply form labels to the left side of the form entry field to which they apply. Our initial source text is in English.



The English looks nice enough. They Malay, on the other hand, looks pretty ugly. The large expansion factor produces unfortunate stacking of the text on the left, and large white spaces to the right. Although the box expands vertically to hold all the text, we are wasting a lot of space and decreasing the amount of information that will appear in the reader's initial screen (you can imagine that this would be compounded by other fixed with boxes on the page).

With the German translation we have a different problem. The long word Benutzeroberfläche doesn't wrap, and so pushes the select boxes beyond the width of the fixed box container. This has the potential to badly affect the layout of other parts of the screen.



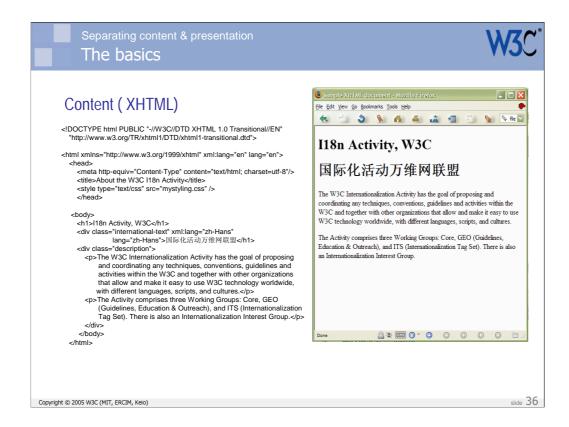
You may want to consider avoiding table cells in such constrained circumstances. This slide shows how the text would look if the input fields were just in a paragraph with the label text.

All the boxes now look fine, and although there is a very slight increase in vertical height overall, we have removed the problems seen with the Malay and German text on the previous slide.

Let's note, again, that this is down to the way the page is **designed/developed**, not the way it is localized. That's a fundamental message of this presentation. Internationalization during design and development removes significant barriers to deploying your content globally.

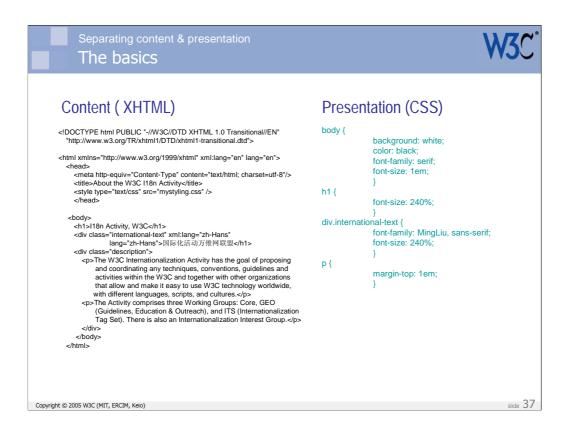


Now we are going to look at the benefits to localization of another good design/development best practice that you would hopefully adopt anyway: the separation of content, presentation and behaviour.



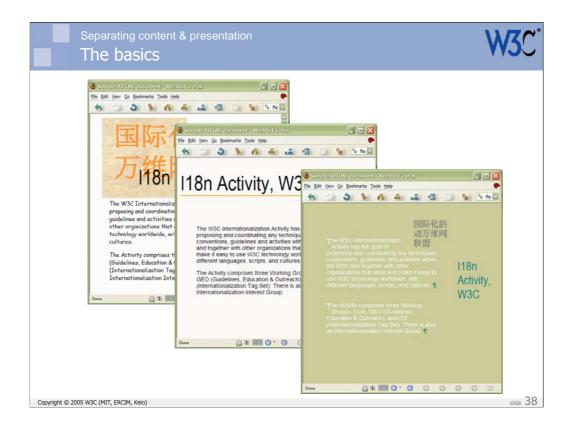
Some HTML is shown on the left. There is no presentational information in the HTML – which is as it should be.

A browser would use default styling to make headers larger and add space between paragraphs.



To make the styling more interesting, you can use a CSS style sheet.

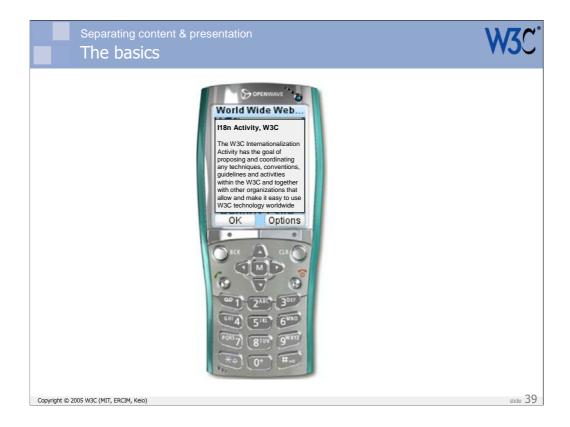
To the right is some CSS code that applies styling to the HTML.



Each of these windows shows EXACTLY the same HTML file. The changes made to the CSS file produced three very different presentations of that basic content.

This is particularly useful for changing the presentational aspects of a site or group of pages. You typically only need to edit a single CSS file, rather than editing all the code of each HTML file.

This can also be beneficial for localization, since typographic approaches, colors, etc, may need to be changed for different locales. Making such changes in the CSS is much easier than adapting the HTML.



Remember, also, that the Mobile Web is becoming increasingly important these days – and may be especially so in developing countries in the future. This means that content needs to be adapted to fit on handheld devices with smaller screens.

Again, this would ideally be achieved by styling the content, rather than writing a completely separate Web.

You should not make assumptions, when creating content, that you know what it will look like when finally displayed. These days, it may well be displayed in a number of different formats.

# Separating content & presentation Style issues for Japanese • problems of resolution to support bold and italics in small CJK characters on-screen • different ways of emphasizing text in Japanese (wakiten & amikake) これは日本語です。 これは日本語です。

Here are some ways in which typographic differences may appear between language versions of the same content. We take the example of Japanese.

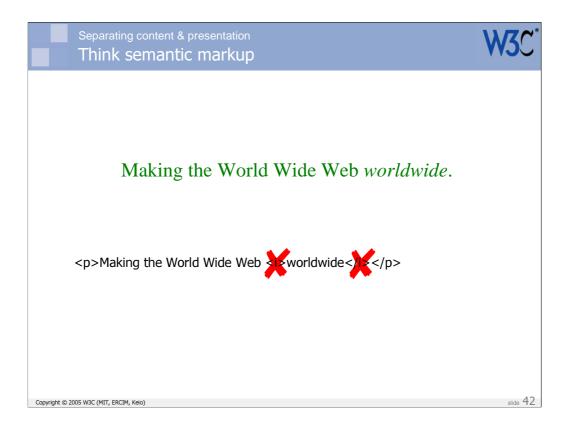
# Style issues for Japanese



- problems of resolution to support bold and italics in small CJK characters on-screen
- different ways of emphasizing text in Japanese (wakiten & amikake)
- different fonts need to be used for font-family assignments
- no upper- vs. lower-case distinction in Japanese
- no convention of distinguishing between proportional and mono-spaced fonts for Japanese scripts
- line heights, paraleading, and possibly font-size may need to be adapted, since Japanese characters are more complicated and square than eg. Latin
- wrapping rules may need adjustment, since there are no spaces between words in Japanese

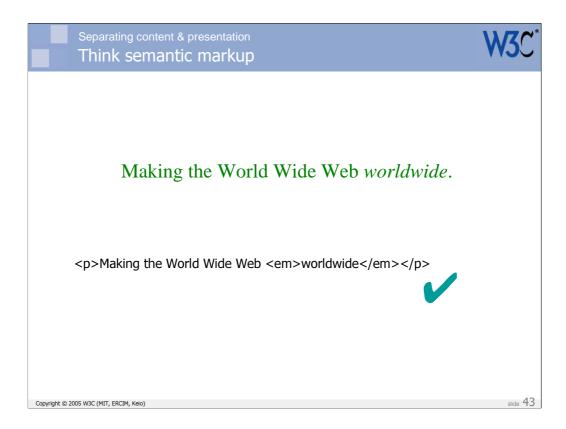
Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 4

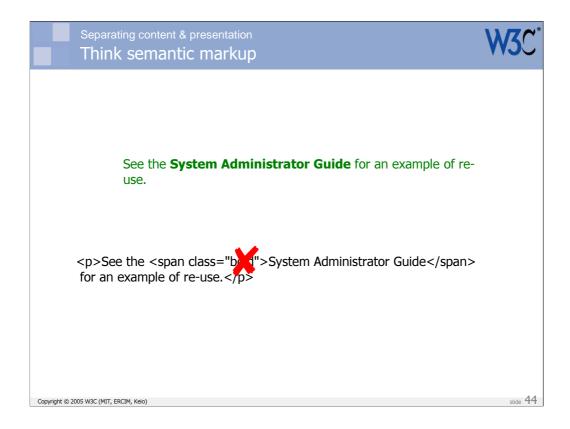


Try to use semantically valid markup where it exists.

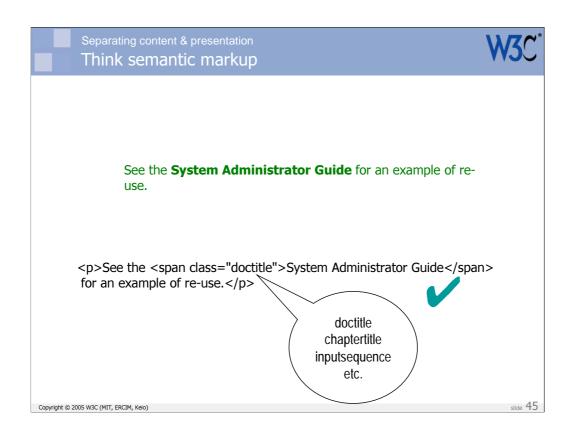
Avoid using the <i> element, use the <em> element instead, if you are wanting to apply a different look because the text is emphasized.

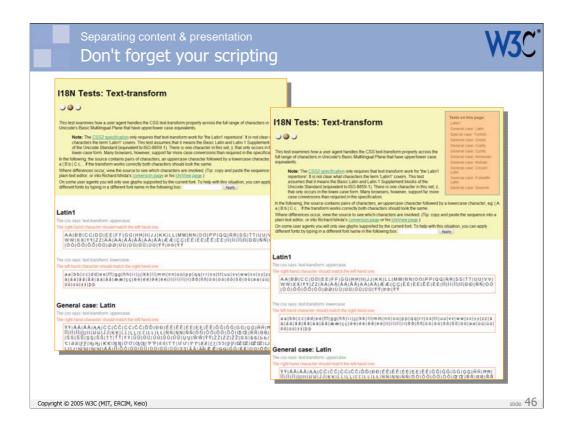


Not only is it better for localization to express the idea or semantics in the content, and leave the presentation to the style sheet, it will also improve your original text by making you more aware of what you are actually doing. (In other words, this is another situation where good authoring practise benefits localization.)



The same applies to use of class names – for example, document conventions such as representation of referenced resources. When using class annotations or microformats, don't describe the expected presentational rendering, describe the function of the text.





You should also consider separation of content and presentation when adding scripting. Let's suppose that we wanted to load some JavaScript after this basic test page has loaded which would automatically add a list of tests on the page to the top right corner. (We may actually want to add links to these tests, but I have resisted that temptation so that the following slides will contain the code examples.)

```
Don't forget your scripting
     function makeList () {
         var headings = document.getElementsByTagName('h2');
         var div = document.createElement('div');
            div.style.cssFloat = 'right';
            div.style.styleFloat = 'right';
            div.style.width = '20%';
             div.style.padding = 'lem';
            div.style.border = 'lpx solid teal';
             div.style.backgroundColor = '#FFCC99';
             div.style.fontSize = '90%';
            div.style.margin = 'lem';
            div.style.color = '#996633;';
         var title = div.appendChild(document.createElement('div'));
             title.style.fontWeight = 'bold';
         title.appendChild(document.createTextNode('Tests on this page:'));
         for (i=0;i<headings.length;i++) {</pre>
             var test = div.appendChild(document.createElement('p'));
             test.appendChild(document.createTextNode(headings[i].childNodes[0].data));
         body = document.getElementsByTagName('body')[0];
         body.insertBefore( div, body.firstChild );
                                                                                             slide 47
Copyright © 2005 W3C (MIT, ERCIM, Keio)
```

Here is a simple function that could be used to add the required text. It creates a div, gets a list of level two headings, and adds the text of the headings to the list.

```
Don't forget your scripting
     function makeList () {
         var headings = document.getElementsByTagName('h2');
         var div = document.createElement('div');
            div.style.cssFloat = 'right';
            div.style.styleFloat = 'right';
            div.style.width = '20%';
            div.style.padding = 'lem';
            div.style.border = 'lpx solid teal';
            div.style.backgroundColor = '#FFCC99';
            div.style.fontSize = '90%';
            div.style.margin = 'lem';
            div.style.color = '#996633;';
        var title = div.appendChild(document.createElement('div'));
             title.style.fontWeight = 'bold';
         title.appendChild(document.createTextNode('Tests on this page:'));
        for (i=0;i<headings.length;i++) {</pre>
            var test = div.appendChild(document.createElement('p'));
             test.appendChild(document.createTextNode(headings[i].childNodes[0].data));
        body = document.getElementsByTagName('body')[0];
        body.insertBefore( div, body.firstChild );
                                                                                             slide 48
Copyright © 2005 W3C (MIT, ERCIM, Keio)
```

Note how we are adding style information directly to the DOM while running this script. This is really obvious in this example, since there is such a lot of it. It is particularly tempting to do this sort of thing if you just want to add a single style effect, such as bolding, to text.

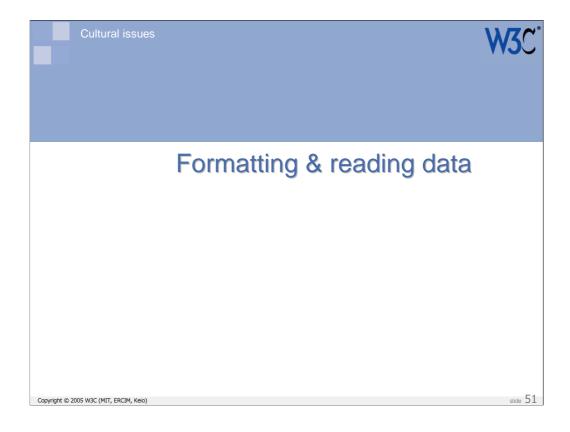
```
Don't forget your scripting
     function makeList () {
         var headings = document.getElementsByTagName('h2');
         var div = document.createElement('div');
             div.setAttribute( 'id', 'testList' );
         var title = div.appendChild(document.createElement('div'));
         title.appendChild(document.createTextNode('Tests on this page:'));
         for (i=0;i<headings.length;i++) {</pre>
             var test = div.appendChild(document.createElement('p'));
             test.appendChild(document.createTextNode(headings[i].childNodes[0].data));
         body = document.getElementsByTagName('body')[0];
                                                               div#testList {
         body.insertBefore( div, body.firstChild );
                                                                          float: right;
                                                                          width: 20%;
                                                                          padding: 1em;
                                                                          border: 1px solid teal;
                                                                          background-color: #FFCC99;
                                                                          font-size: 90%;
                                                                         margin: 1em; color: #996633;
                                                               div#testList div {
                                                                          font-weight: bold;
          see also: Dom Scripting, Jeremy Keith
                                                                                                  slide 49
Copyright © 2005 W3C (MIT, ERCIM, Keio)
```

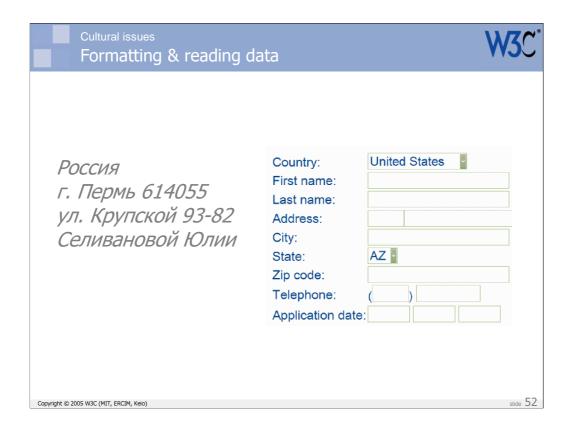
This version of the same function shows a much better approach. We assign an id attribute to the box, then move all the styling information to a CSS file, referencing the markup via the id. This makes the code much cleaner and makes it easier to manage the styling.

Again, this technique is recommended as a standard best practice in Jeremy Keith's book Dom Scripting (which contains many other useful ideas along similar lines). It is another good example of how good web design benefits localization.



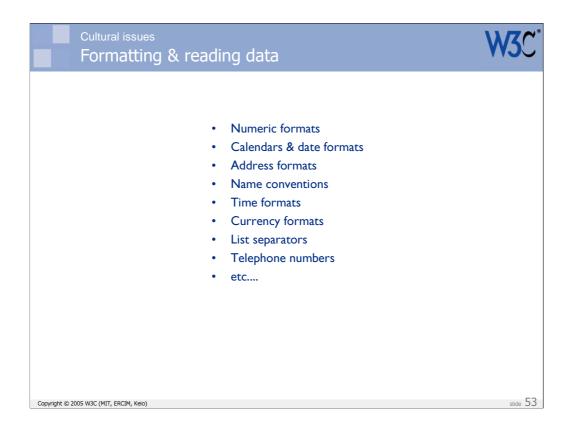
This section will look at a very different set of issues – those related to cultural differences in design.



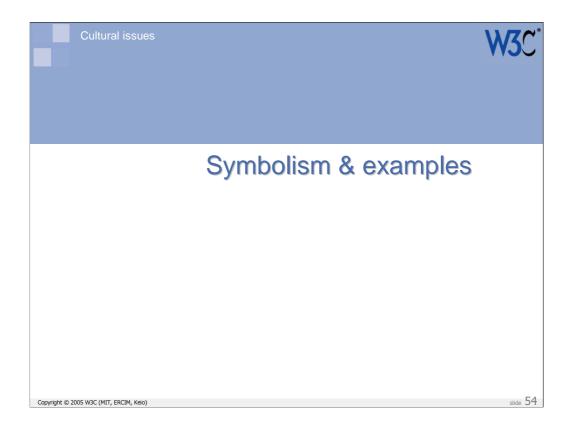


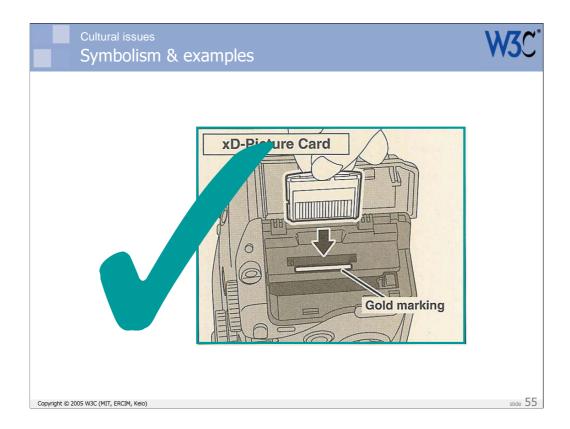
Russian and Japanese addresses are written from the general to the specific, top to bottom. You may need to figure out how to produce these different orderings for forms.

Also, the name of the Russian person above is in the dative case (expressing the idea of 'to the person'). How will you deal with that?



There are a number of ways in which formats differ around the world. Note that recognizing information input into general forms can be more difficult that producing form templates in the right way.

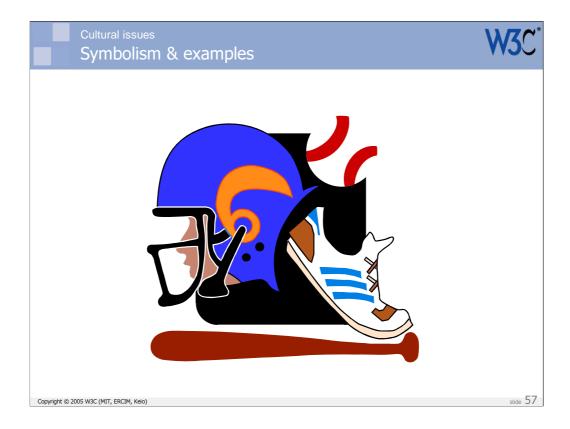




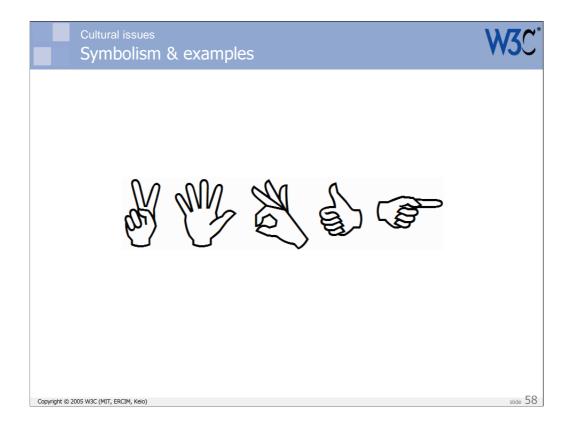
This check symbol means 'correct' or 'ok' in many countries. In some countries, however, such as Japan, it can indicate 'incorrect'. Japanese often convert check marks to circles (their symbol for 'correct') as part of the localization process.



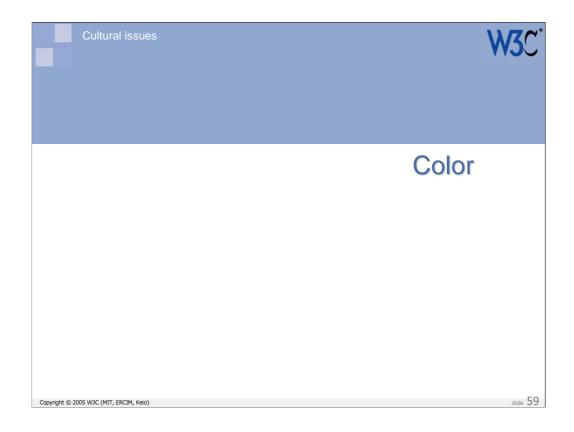
The circles in the columns of this board indicate that space is available, not that there are 0 seats left. It is the equivalent of the check mark.

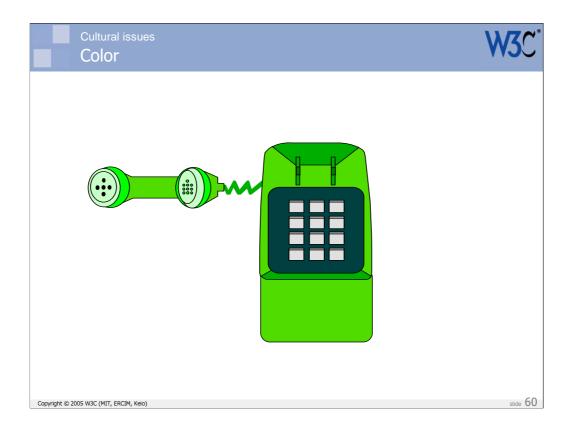


This illustration of sports items is not representative of sports played in the UK, and may need to be changed.

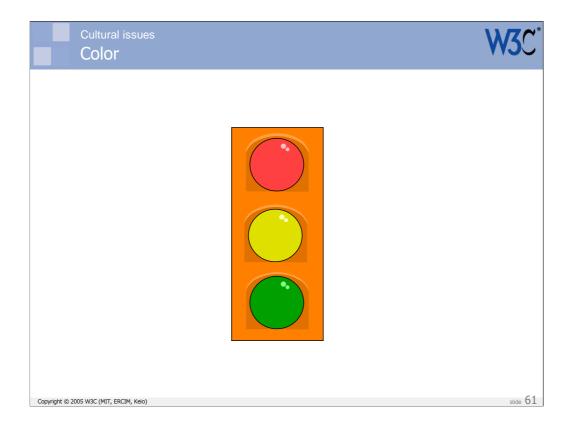


Gestures and sometimes body language can often give completely the wrong message, and should be used with extreme care.

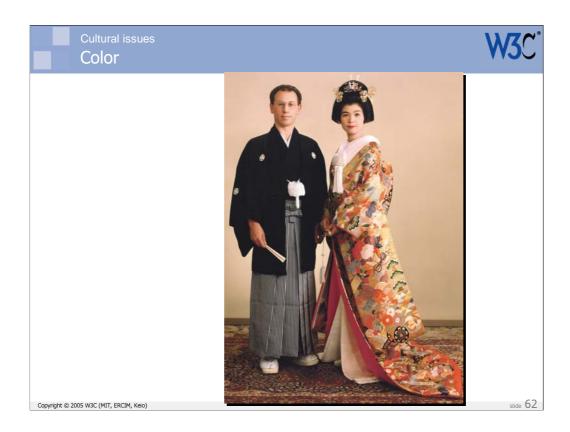




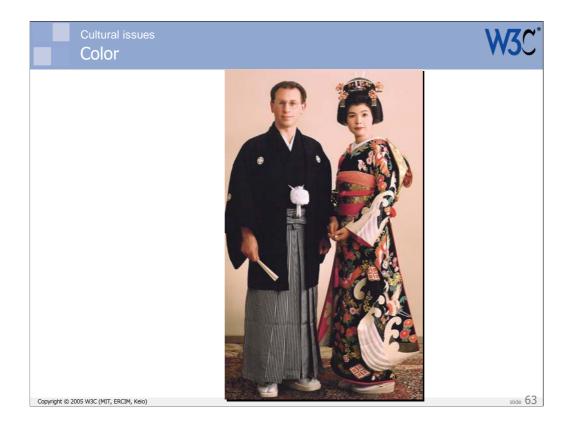
This phone is likely to be perceived immediately as a public telephone in Japan, due to the conventional use of the green color there. In most other parts of the world, this cue is missing. So colors have conventional roles that differ from culture to culture.



Color names also differ from culture to culture, dependent on context. British people often call the middle light here amber, whereas Americans call it yellow. Japanese speaking English will often refer to the bottom light as blue.



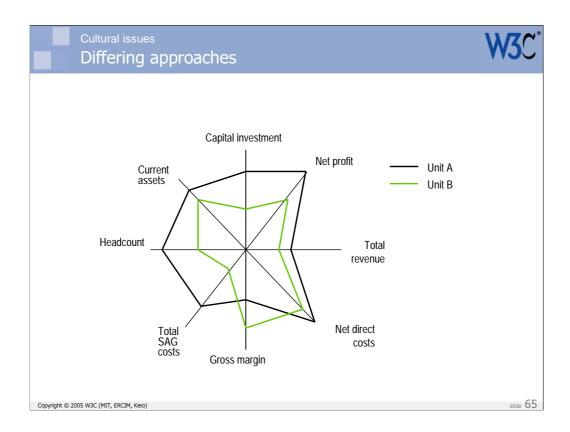
The affective or emotional nature of color can also vary from culture to culture.



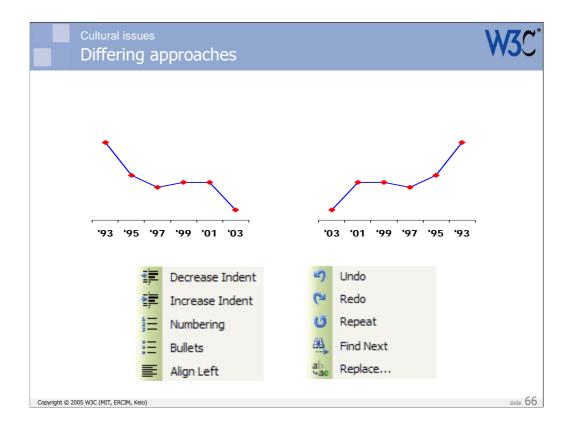
Wearing a black dress for a wedding is not the issue in Japan that it might be in the  $\mbox{\rm UK}.$ 

These differences may also affect the perceptions of your web content in different places around the world.



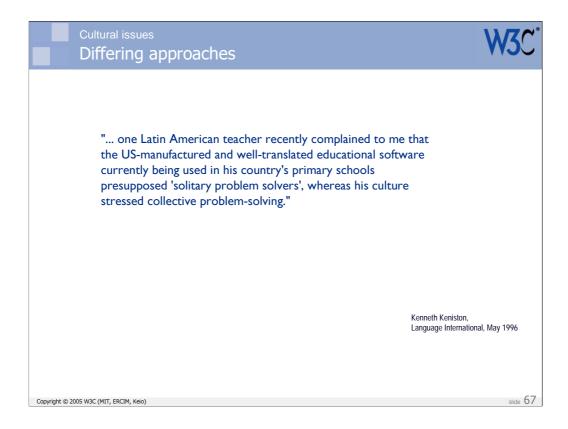


People do things in different ways in different parts of the world. For example, Lotus 1-2-3 was relaunched in Japan with the radar chart after it was discovered that this was a very common way of representing comparative data there.

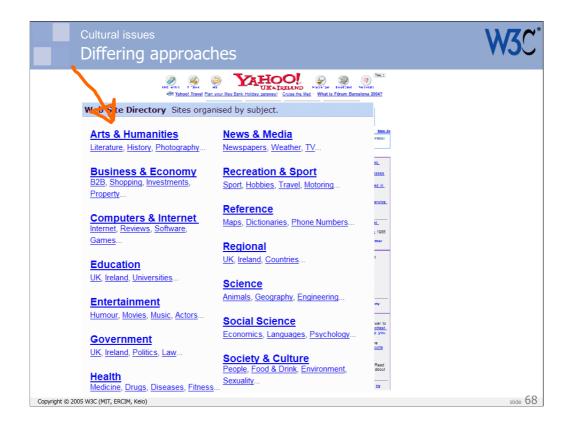


In the Middle East, you may find that tables, spreadsheets, collated pictures and the like need to flow right to left, rather than left to right.

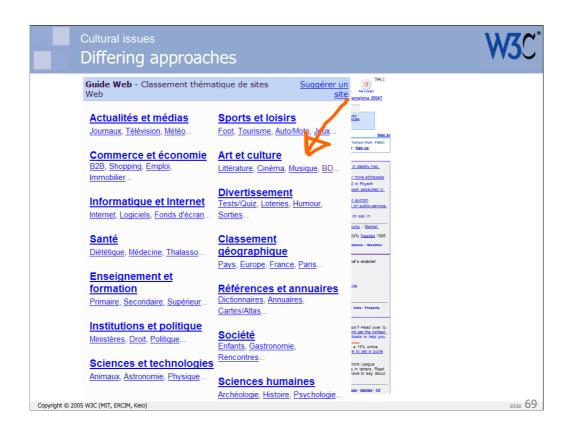
Some graphics with directional bias may need to be mirrored or changed for a predominantly right-to-left context.



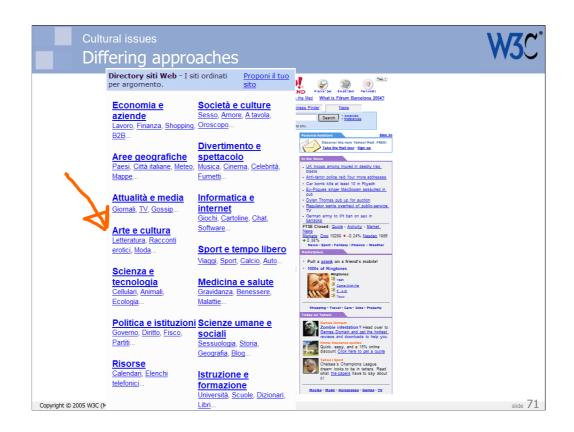
Then there are more fundamental issues about whether the application, product or solution you are developing will actually fit into the foreign culture at all.

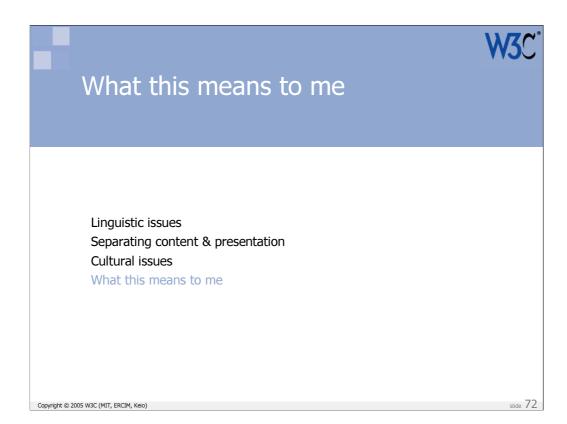


The following slides show how Yahoo adapts the content on its various local home pages, rather than just translating it. This may be something you also need to consider.



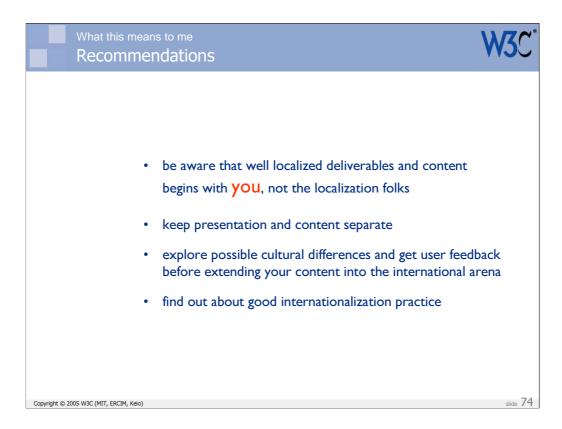




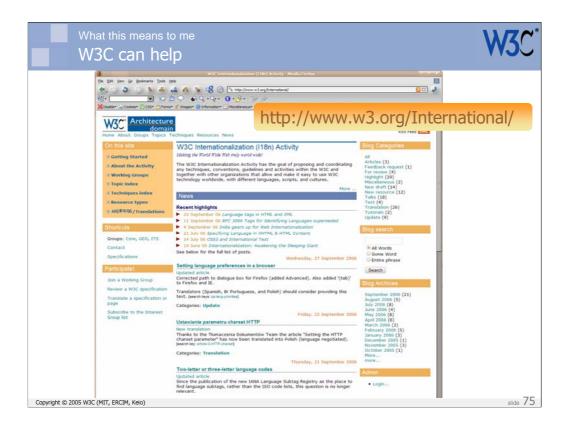


# What this means to me Internationalization means: using a Quality approach to reduce the overall cost and time to market/release of multinational deliverables designing into the deliverable or product an internationalized base, and a modular and easily adaptable architecture not always doing extra work – maybe just working in a better way

Copyright © 2005 W3C (MIT, ERCIM, Keio)



Remember also that, even though you think you don't deal with content that will be internationalized now, you may well need to in the future.



The W3C is trying to provide useful advice at http://www.w3.org/International/

We could always do with help and support for this.

