

INEX 2005 Relevance Assessment Guide

1. Introduction

During the retrieval runs, participating organisations evaluated the 87 INEX 2005 topics (40 content + structure (CO+S) and 47 content-and-structure (CAS) queries) against the IEEE Computer Society document collection and produced a list (or set) of document components (XML elements¹) as their retrieval results for each topic. The top 1500 components in a topic's retrieval results were then submitted to INEX. The submissions received from the different participating groups have now been pooled and redistributed to the participating groups (to the topic authors whenever possible) for relevance assessment. Note that the assessment of a given topic should not be regarded as a group task, but should be provided by one person only (e.g. by the topic author or the assigned assessor).

The aim of this guide is to outline the process of providing relevance assessments for the INEX 2005 test collection. This requires first a definition of relevance (Section 2), followed by details of what (Sections 3) and how (Section 4) to assess. Finally, we describe the on-line relevance assessment system that should be used to record your assessments (Section 5).

2. Relevance in INEX

Relevance in INEX is defined according to the following two dimensions:

- **Exhaustivity (E)**, which describes the extent to which the document component discusses the topic of request.
- **Specificity (S)**, which describes the extent to which the document component focuses on the topic of request.

In order to decrease assessment effort, a highlighting procedure is used in INEX 2005, leading to the following process for assessment (more details in Sections 4 and 5):

- In the first pass, assessors highlight text fragments that contain only relevant information
- In the second pass, assessors judge the exhaustivity level of any elements that have highlighted parts.

As a result of this process, any elements that have been fully highlighted will be automatically labelled as fully specific. The main advantage of this highlighting approach is that assessors will now only have to judge the exhaustivity level of the elements that have highlighted parts (in the second phase). The specificity of any other (partially highlighted) elements will be calculated automatically as some function of the contained relevant and irrelevant content (e.g. in the simplest case as the ratio of relevant content to all content, measured in number of words or characters).

An exhaustivity level is therefore requested for all document components that contain some relevant information, and can be any of the following values:

- ■ **Highly exhaustive (HE)**: the component discusses most or all aspects of the topic of request. In the relevance assessment system, E2 is represented as two green squares.
- □ **Partially exhaustive (PE)**: the component discusses only few aspects of the topic of request. In the relevance assessment system, E1 is represented as two squares, one green and the other white.
- **Too small (TS)**: the component contains some relevant material, but the relevant fragment is too small to be assessed. In the relevance assessment system, TS is a pale green rectangle.

Within the relevance assessment system, a component – that contains some relevant information and has not yet been assessed has an unknown exhaustivity. The corresponding icon is ■ ■ (a green and a red squares). All other elements will be – automatically - assumed as **Not Exhaustive (NE)**.


¹ The terms document component and XML element are used interchangeably.


3. What to judge

Depending on the topic, a pooled result set may contain initially around 500 articles.

Traditionally, in evaluation initiatives for information retrieval, like TREC or CLEF, relevance is judged on document level, which is treated as the atomic unit of retrieval. In XML retrieval, the retrieval results may contain document components of varying granularity, e.g. paragraphs, sections, articles, etc. Therefore, to provide comprehensive relevance assessment for an XML test collection, **it is necessary to obtain assessment for all components at the different levels of granularity that contain any relevant information.**

This means that if you find, say, a section of an article relevant to the topic of the request (i.e. containing highlighted text), you will then need to provide assessment with regards to exhaustivity for the found relevant component, for all its ascendant elements until you reach the article component (unless this can be automatically inferred, e.g. the parent of a highly exhaustive (HE) element will be itself highly exhaustive (HE)), and for all its descendant elements that contain relevant information (i.e. containing highlighted text) until you have identified all relevant sub-components.

Such comprehensive assessments are necessary as it is demonstrated by the following example. Consider the XML structure in Figure 1. Let us say that you judged the marked sec element relevant to the topic, as partially exhaustive (PE, denoted by , see Section 2). Given this single assessment, it would not be possible to deduce the exhaustivity level of the ascending or descending elements. For example, both bdy and article may be judged either highly (HE) or partially exhaustive (PE) depending on the volume of additional relevant information contained within the other sections and in the fm and bm components. Looking at the sub-components of our sec element, it is clear that no conclusions can be drawn from the assessment score assigned to our sec element regarding the exhaustivity level of its sub-components; for example, one of the paragraphs of the second ss2 element may be too small (TS) while the other may be partially exhaustive (PE).

```
[article]
[fm]
...
[bdy]
[  sec]
[ss1]
[ip1]
[ss2]
[p]


[p]


[ss2 ]
[ip1]


[p]


[lc]
- [li]


[p]



[p]

- [li]


[p]


[ss1]
[ss1]
[sec]
...
[bm]
...
```

Figure 1. Example XML structure

As a general rule, it can be said that the exhaustivity level of a parent element is always equal to or greater than the exhaustivity level of its children elements. This is due to the cumulative nature of exhaustiveness. For example, the parent of a highly exhaustive (HE) element will always be highly exhaustive (HE), since the child element already discusses all or most aspects of the topic. However, besides this general rule, no specific rules exist that would automate all the exhaustivity assessment of ascendant and descendant elements of relevant components. Therefore, **you will need to explicitly judge the exhaustivity level of all elements that contain relevant information.** This is the only way to ensure both comprehensive and consistent relevance assessments.

4. How to judge

As described in Section 2, the assessment process is to be done in two phases.

- In the first pass, assessors highlight text fragments that contain only relevant information. A vital consideration is that the highlighting must be based solely on the specificity dimension (e.g. ignoring exhaustivity in the first phase). Assessors should be made aware not to highlight larger contexts because these are more exhaustive, if at the same time they are less specific (i.e. contain irrelevant fragments). It is important that only purely relevant information fragments get highlighted. To decide which text to highlight, you should skim-read the whole article (that a result element is a part of - even if the result element itself is not relevant!) and identify any relevant information as you go along. The on-line system can assist you in this task by highlighting keywords (that are chosen using the interface) and pool elements (elements retrieved by participating systems) within the article (see Section 5).
- In the second phase, you should assess the exhaustivity of the components that intersect with any of the highlighted passages (i.e. identified in the first phase). The on-line assessment system (see Section 5) will identify for you all elements that have to be assessed for phases 2.

During the relevance assessment of a given topic, all parts of the topic specification should be consulted in the following order of priority: narrative, topic description, and topic title. The narrative should be treated **as the most authoritative description of the user's information need**, and hence it serves as the main point of reference against which relevance should be assessed. In case there is conflicting information between the narrative and other parts of a topic, the information contained in the narrative is decisive. Note that it is not because that a term listed within the topic is not present in an element that the element is not relevant. It may be that a component contains some or maybe all the terms, but is irrelevant to the topic of the request. Also, there may be components that contain none of the terms yet are relevant to the topic.

For both the CO+S and CAS topics, the topic titles (may) contain structural constraints in the form of XPath expressions. These structural conditions should be ignored during your assessment. This means that you should assess the elements returned for a CO+S and CAS topic as whether they satisfy your information need (as specified by the topic) **with respect to the content criterion only**.

Note that some result elements may be related to each other (ascendant/descendant), e.g. an article and some sections or paragraphs within the article. This should not influence your assessment. For example if the pooled result contains Chapter 1 and then Section 1.3, you should not assume that Section 1.3 is more relevant than Sections 1.1, 1.2, and 1.4, or that Chapter 1 is more relevant than Section 1.3 or vice versa. Remember that the pooled results are the product of different retrieval engines, which warrants no assumptions about the level of relevance based on the number of retrieved related components!

You should judge each document component on its own merits! That is, a document component is still relevant even if it the twentieth you have seen with the same information! It is imperative that you maintain consistency in your judgement during assessment. Referring to the topic text from time to time will help you maintain judgement consistency.

5. Using the on-line assessment system (X-Rai)

There is an on-line relevance assessment system (XML Retrieval Assessment Interface) provided at:

<https://inex.lip6.fr/2005/xrai>

which allows you to view the pooled result set of the topics assigned to you for assessment, to browse the IEEE-CS document collection and to record your assessments. Use your INEX username and password to access this system.

The assessment tool works with opera and recent "gecko" browsers: we highly recommend you to use Opera (version 8 or up only) available at <http://www.opera.com>. Other compatible browsers are:

- **Mozilla** (version 1.7 or up) at <http://www.mozilla.org/products/firefox/>.
- **Firefox** (version 1 and up) at <http://www.mozilla.org/products/mozilla1.x/>.

Note that **JavaScript must be enabled** for the assessment tool to work and that **the assessment tool is not compatible with Internet Explorer**. Any bug report should be submitted using the project homepage (<https://developer.berlios.de/projects/x-rai/>) using the link in the “Links” menu of the interface (Figure 2).

5.1. Home page

After logging in, you will be presented with the Home page (see Figure 2) listing the topic ID numbers of the topics assigned to you for assessment (under the title “Choose a pool”). This page can always be reached by clicking on the “X-Rai” link of the menu bar on any subsequent pages.








Figure 2: Home page and menu bar

In the “Links” menu

- **INEX 2005:** link to the official INEX web site.
- **X-Rai project:** link to the development web site of X-Rai where you can submit bug reports or/and feature requests.
- **Guide:** the latest version of this assessment guide.

Each X-Rai page is composed of the following components:

- The menu bar, which is itself composed of four parts:
 1. The login name (e.g. “demo” in Figure 2),
 2. A list of menu items, which can be accessed by holding the mouse over the menu label (e.g. “**Links**” in Figure 2.),
 3. The location within X-Rai, where each location step is a hyperlink (in Figure 2, we are at the root of the web site, so the only component of the location is “**X-Rai**”, which is a link to the home page),
 4. The menu bar may also contain a number of icons (displayed on the right hand side, see Figure 3a). Click on one of these icons to display (or hide):
 -  Information about X-Rai.
 -  Toggle the help
- The main window.
- An optional status bar (see Figure 5), displayed only when assessing a pool, i.e. in pool, sub-collection or article view (see relevant sections below) appears at the bottom of the window and shows the number of unknown assessments you have to judge before completing assessing the document (in Figure 5, there is only one unknown assessment).

- In the status bar, three arrows (,  and ) may be used to navigate quickly between the elements to be assessed. You may also use the shortcut keys of 1 (left), 2 (up) and 3 (right). The up arrow enables you to move to a level up in the hierarchy, e.g. from an article or a collection part to its innermost enclosing part of the collection (you move in the opposite direction by selecting a sub-collection or an article). The left arrow can be used to go to the previous element to be assessed, while the right arrow to go to the next element to be assessed.

The on-line assessment system provides three main views (Sections 5.2 to 5.4):

1. Pool view,
2. Sub-collection view, and
3. Article view

5.2. Pool view

Clicking on a topic ID will display the Pool main page for that topic (see Figure 3a).

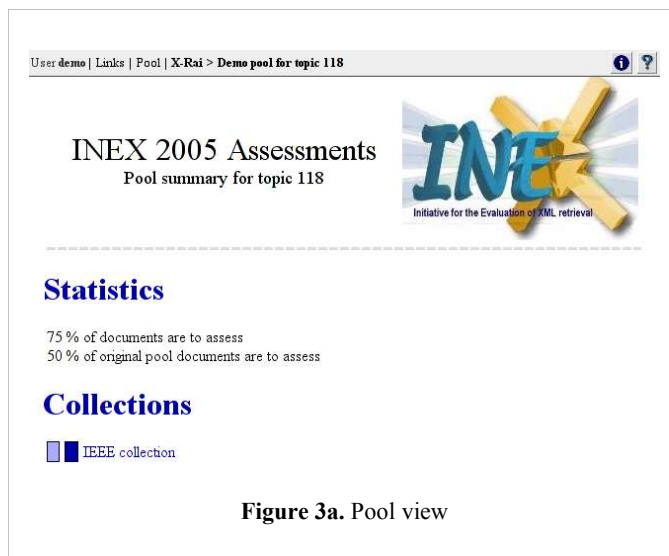


Figure 3a. Pool view

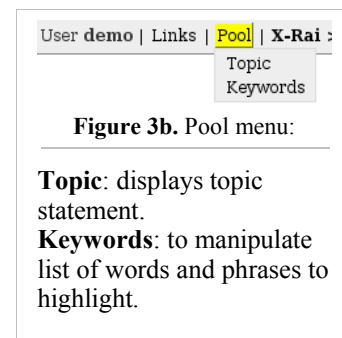


Figure 3b. Pool menu:

Topic: displays topic statement.
Keywords: to manipulate list of words and phrases to highlight.

Here, a new menu item, “**Pool**”, appears on the menu bar at the top of the window.

Within the “**Pool**” menu (Figure 3b), with the “**Topic**” submenu item you can display the topic statement in a popup window. This is useful as it allows you to refer to the topic text at any time during your assessment.

The “**Keywords**” submenu item allows you to access a feature, where you can specify a list of words or phrases to be highlighted when viewing the contents of an article during assessment. These cue words or phrases can help you in locating potentially relevant texts within an article and may aid you in speeding up your assessment (so add as many relevant cue words as you can think of!). You may edit, add to or delete from your list of keywords at any time during your assessment (remember, however, to refresh the currently assessed article to reflect the changes). You may also specify the preferred highlighting colour for each and every keyword. After selecting the “**Keywords**” menu item, a popup window will appear showing a table of coloured cells. A border surrounding a cell signifies a colour that is already used for highlighting some keywords. Move the mouse over a coloured cell to display the list of keywords that will be highlighted in that colour. To edit the list of words or phrases for a given colour, click on the cell of your choice. You will be prompted to enter a list of words or phrases (one per line) to highlight. You can choose three different highlighting modes using the drop-down menu: using coloured fonts, drawing a border around the phrase or using a background colour. Note that the words or phrases you specify will be matched against the text in the assessed documents in their exact form, *i.e.* no stemming is performed.

Under the title “**Collections**” is the list of collections to be assessed. In INEX 2005 (ad hoc task) there is only one such collection, the IEEE collection.

The left or right arrows on the status bar move the focus to the previous or next collection, where there is at least one element to assess (since there is only one collection, so no change will occur).

Clicking the hyperlink of “IEEE collections” will take you into the sub-collection view.

5.3. Sub-collection view

The sub-collection views allow you to browse the different sub-collections within the IEEE collection, i.e. volumes, years within a given volume (see Figure 4), the collection of articles within a given volume and year. Note that this view will show all elements within a sub-collection, i.e. all articles within a given volume and year, and not only the ones that need to be assessed. For each possible sub-collection, there is an indication on the number of documents to be assessed in it (if this number is greater than 0), both for documents that were initially in the pool and for documents you choose to assess.

The left or right arrows on the status bar move the focus to the previous or next sub-collection, where there is at least one document to assess. You can also directly click on a link to a sub-collection.

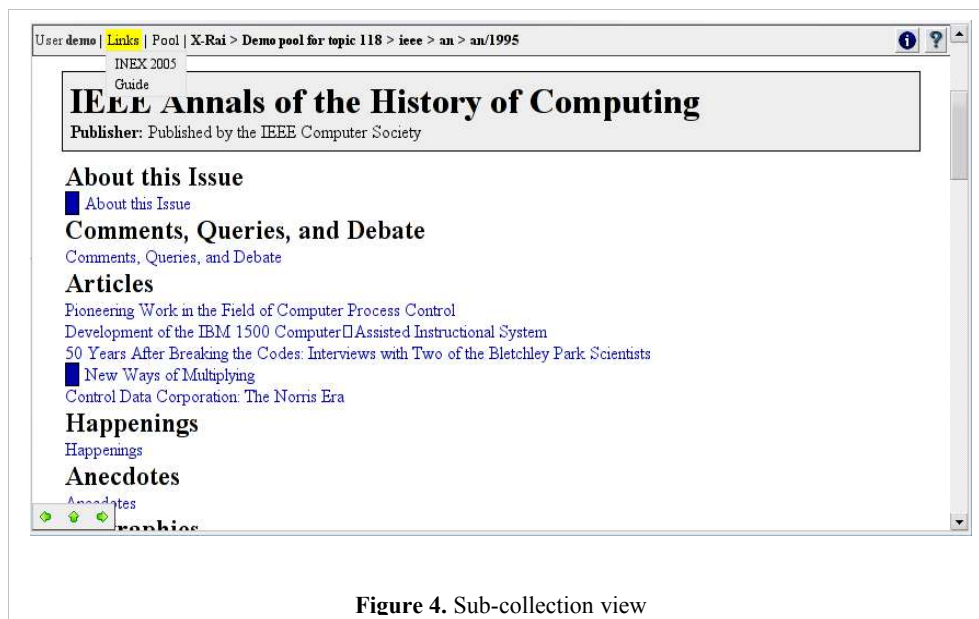


Figure 4. Sub-collection view

5.4. Article view

It is in this article view that elements can be assessed. The article view (see Figure 5) displays all the XML elements of an article together with their content. There are two types of objects within an article view: XML elements and passages. The latter are defined by the assessor while highlighting whereas the former are predefined by the XML file. XML elements boundaries are denoted by < and > (less/greater in pale blue). **A passage in the interface has a yellow background and is enclosed within two braces on a blue background like this sentence**. For each element, the assessment value is displayed at its start, that is after the “<” for an XML element.

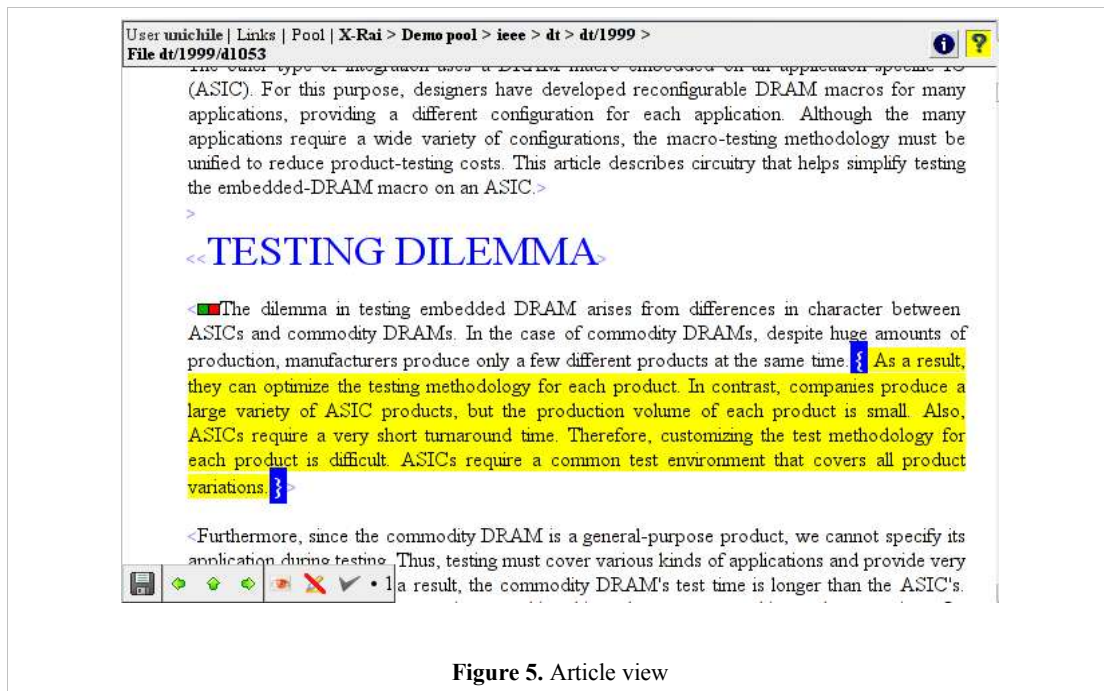



Figure 5. Article view


Highlighting

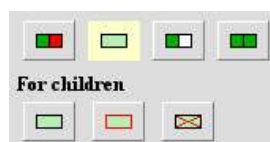
You are in the highlighting mode when the marker icon in the status bar is . During the highlight phase, you should identify only relevant (i.e. totally specific) passages by highlighting them. Passages can span over XML element boundaries. The passage limits are predefined by a pre-processing of XML files and correspond “more or less” to sentence boundaries. A consequence of this is that you should highlight the smallest passage that encloses the only relevant information if the predefined boundaries do not correspond exactly to the totally specific fragment.

To highlight a passage, select it with the mouse as you would do in any word processor or text editor, and click on the square with the yellow background (or press “h”).

If you make an error, you can unhighlight it by selecting the non relevant passage and clicking on the square with the white background (or press “u”).

Assessing

Once you have finished to highlight relevant passages, you may switch to the assessing mode by clicking on the yellow marker. You can also switch by pressing the “m” (for **mode**) key. Once you have switched, it is not possible to (un)highlight passages any more and the icon in the status bar should look like a crossed marker . To assess an element, simply click on the assessment icon. An assessment panel will then appear:



The first line shows the list of possible assessments for the current assessed element. The second line contains shortcuts to set or remove “too small” judgements for some or all of the children - and, as a consequence, to all the descendants: the first button assesses **all the non judged** children as “too small”; the second one assesses **all** children as “too small”, and the third one resets the


assessments to “unknown” for children judged “too small”. The process is recursive: if a value is changed for a child, then the same operation is applied for its own children. For example, judging a child as “too small” will judge the child children as TS, etc.



You should always give the exhaustivity value of the **strictly** contained data, that is without taking into account the component context (i.e. parent or sibling component). Any XML element that contains at least a passage cannot be judged as “too small”. The interface will prevent you from doing so (in the above Figure, the “too small” assessment is disabled).

In order to assess an XML element, you can have a quick look at its boundaries by putting the mouse pointer over an element assessment: the content of the XML element will be bordered by a red line. You can also change the background colour of an XML element by clicking on its assessment while pressing the shift key. You can switch back to the normal display by clicking again on its assessment while pressing the shift key. This is useful when assessing big elements (like a section, etc.) so you can inspect their full content before judging them.

Judging an element (or its descendants) “too small”

If the assessed XML element intersects with one or more passages but does not contain any one passage completely, it is possible to assess it as “too small”. This will also automatically assess its descendants as TS. When the element - denoted X in the following - is not too small itself but its descendants (or a part of them) are, then it is possible to judge all of them as “too small”. The procedure to follow is:

1. assess explicitly all the descendants which are not “too small”
2. click on the X assessment icon, and press on the icon “assess the remaining children as too small” (the icon ). Note that the interface will automatically judge as TS any descendant of a too small element.

It is also possible to judge all the descendants as “too small”, overriding their values by clicking on the TS icon with a red border (). If you made a mistake and want to reset the assessments of the descendants, you can also remove all the TS judgements of the descendants by clicking on the crossed pale rectangle icon (.

Assessment consistency

Contrarily to last years, there is no pre-checking on the allowed exhaustivity values. If there is a conflict after the user assessed an element, then the conflicting assessment value(s) is/are reset to “unknown” so the judge has to reassess it: for example, let a section be assessed as HE and its only paragraph as HE. If the user changes the paragraph assessment to PE or unknown/too small, then the section assessment is set to unknown.

Another example of consistency check is when you change an assessment from “too small” to another value: the descendants, which were previously assessed as “too small”, will be reset to “unknown”.

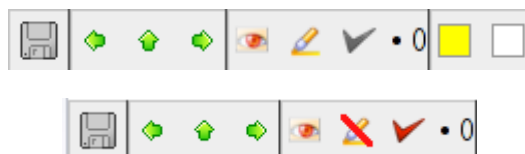


Figure 6. Status bar (article view only): highlighting mode (top) and assessing mode (bottom)

The disk icon (here disabled): saving your assessments

The left/right arrows: going to the previous/next element to judge

The up arrow: going to the sub-collection view that contains the article

The eye: shows or hides the pool elements

The yellow marker: switch between the highlighting and the assessing view.

The mark reflects the status of the document: completely assessed and validated (green), completely assessed but not validated (red), and not completely assessed and not validated (grey). You can validate a document (i.e., mark it as finished) only if the mark is red.

The number indicates the number of elements to be assessed.

The yellow/white square (when in highlighting mode) permits to (un)highlight the selected passage.















5.6. Saving your assessments

The assessment tool this year does not automatically save the assessments, but you NEED TO SAVE YOUR RELEVANCE ASSESSMENTS by clicking on the disk icon:



The icon is disabled (grey shade) when all assessments are saved.

Be warned that Opera doesn't provide a way to prevent from exiting a page without saving assessments. PLEASE ONLY USE THE INTERFACE TO NAVIGATE INTO THE SITE as this is the only way to prevent you from leaving a page with non-saved assessment(s).

<i>Icon</i>	<i>Shortcut</i>	<i>Action description</i>
All views within a pool		
	1	Highlight the previous element, (sub)collection or document to assess
	2	Go to the container (sub-collection for an article, etc.)
	3	Highlight the next element, (sub)collection or document to assess
Article view		
	control+s	Save the current assessment
		Hide the pool elements
		Show the pool elements
Article view - highlighting mode 		
	h	Highlight the currently selected passage.
	u	Unhighlight the currently selected passage
	m	Finish highlighting and switch to the assessing mode
Article view - assessing mode 		
		Mark the article as finished
		Mark the article as not finished
	m	Go back to the highlighting mode

September 2005
Mounia Lalmas and Benjamin Piwowarski