

INEX 2004 Relevance Assessment Guide

1. Introduction

During the retrieval runs, participating organisations evaluated the 74 INEX 2004 topics (40 content-only (CO) and 34 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 2004 test collection. This requires first a definition of relevance for XML retrieval (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 dimensions: exhaustivity and specificity

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.

Exhaustivity is measured on the following 4-point scale:

Not exhaustive (E0): the document component does not discuss the topic of request at all.

Marginally exhaustive (E1): the document component discusses only few aspects of the topic of request.

Fairly exhaustive (E2): the document component discusses many aspects of the topic of request.

Highly exhaustive (E3): the document component discusses most or all aspects of the topic of request.

Specificity is assessed on the following 4-point scale:

Not specific (S0): the topic of request is not a theme of the document component.

Marginally specific (S1): the topic of request is a minor theme of the document component (i.e. the component focuses on other, non-relevant topic(s), but contains some relevant information).

Fairly specific (S2): the topic of request is a major theme of the document component (i.e. the component contains mostly relevant content and only some irrelevant content).

Highly specific (S3): the topic of request is the only theme of the document component.

Although the two dimensions are largely independent of each other, a not-exhaustive (E0) component can only be not specific (S0) and vice versa. Other than this rule, a component may be assigned any other combination of exhaustivity and specificity, i.e. E3S3, E3S2, E3S1, E2S3, E2S2, E2S1, E1S3, E1S2, and E1S1. For example, a component assessed as E1S1 is one that contains only marginally exhaustive relevant information (E1) where this relevant content is only a minor theme of the component, i.e. most of the content is irrelevant to the topic of request (S1).


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 between 500 and 1,500 document components of 500 - 510 articles, where a component may be a title, paragraph, subsection, section, or whole article, etc.

Traditionally, in evaluation initiatives for information retrieval, like TREC, 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, you will then need to provide assessment –(both with regards to exhaustivity and specificity) for the found relevant component, for all its ascendant elements until you reach the article component, and for all its descendant elements 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, which encapsulates all text fragments relevant to the topic, as highly exhaustive and fairly specific (E3S2, denoted by , see Table 1). Given this single assessment, it would not be possible to deduce the exhaustivity and specificity levels of the ascending or descending elements. For example, both `bdy` and `article` may be judged either fairly (S2) or marginally specific (S1) depending on the volume of additional, irrelevant 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 or specificity levels of its sub-components; i.e., any of the `ss1`, `ss2` subsections, and `p` paragraphs (etc.) may be highly (E3), fairly (E2) or only marginally (E1) exhaustive, and could be highly (S3), fairly (S2) or only marginally (S1) specific, or could even be irrelevant (E0S0). For example, one of the paragraphs of the first `ss2` element may be irrelevant (E0S0), while the other may be fairly exhaustive and highly specific (E2S3).


```
[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 (E3) element will always be highly exhaustive (E3), since the child element already discusses all or most aspects of the topic. Another rule for the exhaustivity dimension is that a component whose child elements are all not exhaustive (E0)

will also be not exhaustive (E0). A rule regarding specificity is that the parent of an element whose specificity degree is greater than 0, must also have a specificity level greater than 0, but less or equal to the maximum S value of all its child elements. For instance, suppose that a parent element has a small child element with S1 and a large child element with S2, then the S value of that parent can only be either 1 or 2. However, besides these general rules, no specific rules exist that would automate all the assessment of ascendant and descendant elements of relevant components. Therefore, **you will need to explicitly judge all elements that contain relevant information**. This is the only way to ensure both comprehensive and consistent relevance assessments.

4. How to judge

To assess the exhaustivity and specificity of document components, we recommend a three-phase approach.

- During the first phase, 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 will assist you in this task by highlighting keywords within the article (see Section 5).
- In the second phase, you should assess the exhaustivity and specificity of the relevant components (i.e. identified in the first phase), and that of their ascendant and descendant XML elements.
- To ensure comprehensive assessments, in the third phase, you should assess the exhaustivity and specificity of the descendant XML elements of all elements that have been assessed as relevant during the second phase.

The on-line assessment system (see Section 5) will identify for you all elements that have to be assessed for phases 2 and 3.

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, topic title and keywords. 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. The keywords should be used strictly as a source of possibly relevant cue words and hence only as a means of aiding your assessment. You should **not rely only on the presence or absence of these keywords** in document components to judge their relevance. It may be that a component contains some or maybe all the keywords, but is irrelevant to the topic of the request. Also, there may be components that contain none of the keywords yet are relevant to the topic. The same applies to the terms listed within the topic title!

In the case of CAS topics, the topic titles contain structural constraints in the form of XPath expressions. Since these structural conditions are there to provide hints for the search engines, they should be ignored during your assessment. This means that you should assess the elements returned for a CAS topic as whether they satisfy your information need (as specified by the topic) **with respect to the content criterion only**. Therefore, you should not assess an element as “irrelevant” only because the structural condition is not satisfied.

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:

<http://inex.lip6.fr/2004/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 recent "gecko" browsers: we highly recommend you to use either

- **Mozilla** (version 1.7 or up) at <http://www.mozilla.org/products/firefox/>.
- **Firefox** (version 0.9.1 or 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**.

X-Rai uses MathML in order to render mathematical formulas. Some specific fonts have to be downloaded if you want a perfect rendering of mathematical formulas. More details can be found at the following address: <http://www.mozilla.org/projects/mathml/fonts/>.





Before detailing the assessment system, first we define describe the graphical scheme employed to represent the possible combinations of relevance values.

5.1. Relevance values

In the on-line assessment system, the following scheme is used for representing assessment values (see also Table 1):

- **Exhaustivity** level is displayed in different shades of blue, where the darker the blue, the more exhaustive the element.
- **Specificity** level is represented by two overlapping discs: the filled in disc denoting the relevant content and the white disc representing the irrelevant content. A single blue disk represents a highly specific (S3) element; a blue disk in front of the white disc denotes a fairly specific (S2) element; a blue disk behind the white disc corresponds to the marginally specific (S1) degree; and finally a single white disc represents a not-specific (S0) component.

The tables below show the different icons used to indicate the relevance value of an XML element.

	Element is not assessed
	Element is to be assessed
	Element is irrelevant (E0S0)
	Element is inconsistent












Exhaustivity Specificity	Highly exhaustive (E3)	Fairly exhaustive (E2)	Marginally exhaustive (E1)
Highly specific (S3)			
Fairly specific (S2)			
Marginally specific (S1)			

Table 1. Icons used to indicate relevance values

Note that all icons except the  icon can be used by assessors to specify the relevance value (the exhaustivity and specificity level) of an element. The  icon is used by the on-line assessment system only to mark components that are in an inconsistent state.

5.2. Home page

After logging in, you will be presented with the Home page (see Figure 2a) 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.

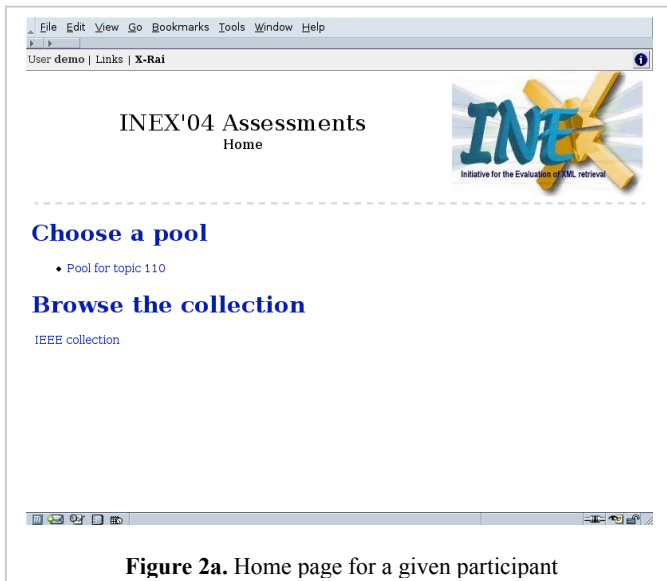


Figure 2a. Home page for a given participant

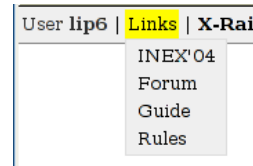


Figure 2b. Links menu:

- **INEX'04**: link to the official INEX web site.
- **Forum**: link to the site, where discussions and bug reports related to the assessment rules and system can be read and posted.
- **Guide**: the latest version of this assessment guide.
- **Rules**: a technical report on the passive and active inference rules used in X-Rai (see Inference Rules section).

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 2a and “lip6” in Figure 2b),
 2. A list of menu items, which can be accessed by holding the mouse over the menu label (e.g. “**Links**” in Figure 2b.),
 3. The location within X-Rai, where each location step is a hyperlink (in Figure 2a, 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.






The tree view of the article (only available in article view).



The current list of bookmarks (only available in article view).

- The main window.
- An optional status bar (see Figure 3a and 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 statistics on the current view for each relevance value, e.g. how many elements have been assessed as highly exhaustive and highly specific (), as highly exhaustive and fairly specific (), etc; as irrelevant (); and how many elements remain to be assessed (). Only when no more elements remain to be assessed is the assessment for that view (pool / sub-collection / article) complete.

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 **shift + left / up / down**. 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:

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

5.3. Pool view

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

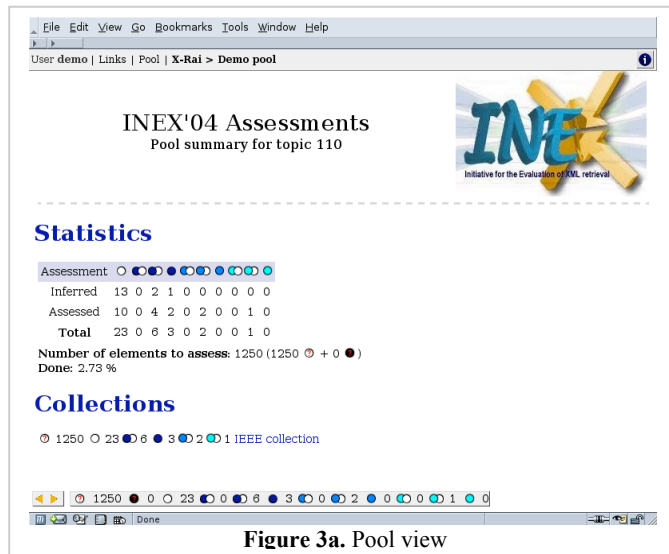


Figure 3a. Pool view

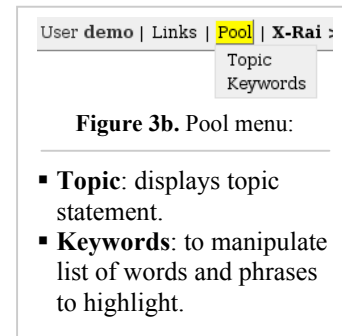


Figure 3b. Pool menu:

- **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’04 (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, the focus will remain).

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

5.4. 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.

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 element to assess. You can also directly click on a link to a sub-collection.

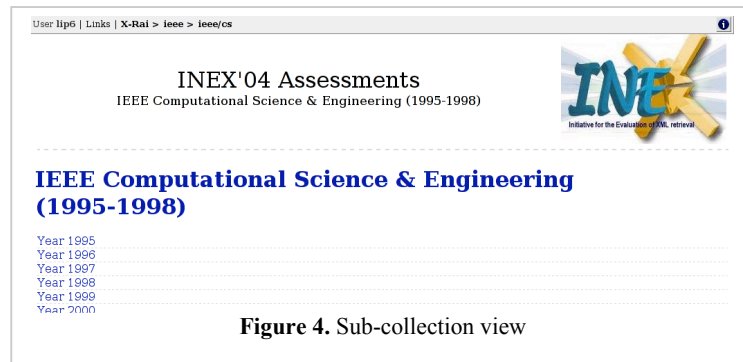


Figure 4. Sub-collection view

5.5. Article view

It is in this article view that elements can be assessed. The article view (see Figure 5) displays all components of an article, whether these elements are to be assessed or not. In addition, the article view shows every XML tag in the article while keeping an eye-friendly view of the article. XML tags are displayed between brackets, in light blue font. For each component the currently assigned (or inferred) relevance values are displayed in front of the XML tag name.

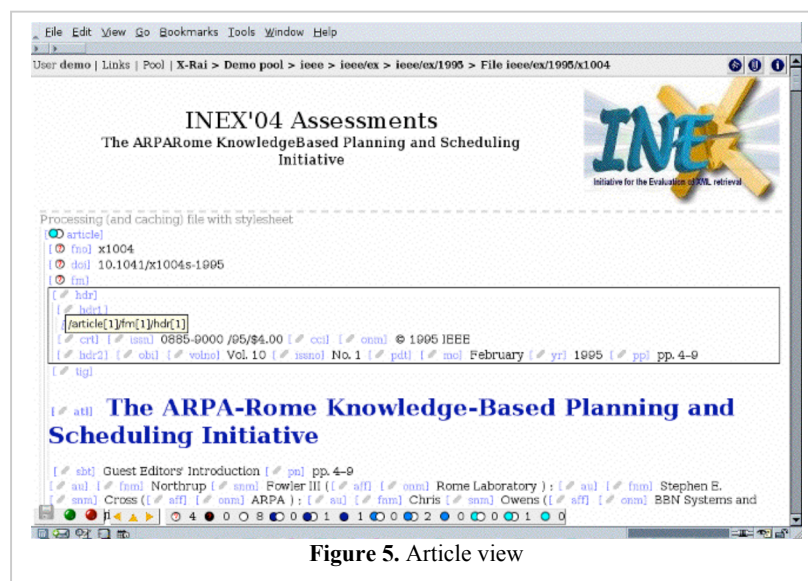




Figure 5. Article view

For instance, an `abs` element, which has been assessed as highly exhaustive and highly specific (E3S3), is displayed with the following XML tag syntax:




[ `abs`]




To make an assessment, first hold the mouse over an XML tag name. The cursor will turn into a cross shape. You can then:

- Left-click to display the assessment panel for the element. The assessment panel has three components: the XPath of the selected element (first line), the current assessment value (second line), and the set of 11 icons (reflecting all possible assignments as shown in Table 1). Forbidden assessments (e.g. assessing a parent element as not relevant where one of its child elements is relevant) are displayed in a grey box. To assess the current element, click on the icon with the corresponding relevance value. To hide the panel, click anywhere else in the panel.

- It is possible to assess groups of elements. Control-click to select the element or control-double-click to select the sibling elements in the same “state” (i.e. elements which are assessed or un-assessed). Then click on the green button  or press the key **shift-G** to display an assessment panel for the selected elements. Click on the red button  or press the key **control-shift-G** to clear the current selection.
- Right-click to display the navigation panel (see figure below). Depending on the element you clicked on, there might be up to three arrows. Click on the left (or right) arrow to access the previous (or next) sibling element in the article tree. Click on the up arrow to access the parent. If necessary, the window scrolls up or down to make the element tag visible. The element tag is then highlighted in red for a brief moment.



The  icon at the bottom centre of the navigation panel can be used to add the currently selected element to the list of bookmarks. Click on  to remove the element from the bookmarks. To display the list of bookmarks, click on the clip icon  or press the key **B**. The bookmarks are ordered with respect to the article they occur in. Click on a bookmark to highlight the element.

It is also possible to use a document-tree view in order to navigate into the article. Click on the icon  or press the key **T**. A panel then appears with the document tree view. Click on  to expand a sub-tree or on  to collapse a sub-tree. Click on any element to highlight it in the article view.



5.6. Saving your assessments

Contrary to last year, 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.

5.7. Inference rules

This year, the assessment system makes use of two types of inference mechanisms to ensure exhaustive and consistent assessments: we refer to these as passive and active inferences. The passive type simply identifies new elements to be assessed based on those already assessed. For example, for any relevant element (e.g. any component assessed other than “irrelevant”), the relevance of its child elements must be assessed, even if these were not part of the original assessment pool (i.e. have not been retrieved). With the application of the passive inference rules, these need-to-be-assessed components will be marked with the  icon. Unlike the passive rules, the active inference rules are able to derive the relevance value of some elements. These inferred relevance values will be marked using a red border. For example,  denotes “inferred as not relevant”, which is (for example) assigned to a component if all its child elements have been assessed as “not relevant”.