



# INEX 2009 Book Track

## Topic Development Guidelines

---

Gabriella Kazai, Marijn Koolen, Monica Landoni

Version 2.0

### 1 Introduction

Both the *Book Retrieval* and the *Focused Book Search* tasks aim at evaluating IR/XML IR approaches over a collection of digitized books. In order to enable the evaluation of the various relevance ranking strategies, the Book Track is building a test collection based on a corpus of over 50,000 books, and comprising a set of user topics, and relevance assessments.

*Topics are representations of users' information needs and may comprise of several aspects or sub-topics. Relevance assessments are sets of books and book parts relevant to a topic or an aspect of a topic.*

Both the topics and relevance assessments are to be created by the participating organisations. Thus, each participant plays a vital role in building the test collection. The quality of the resulting collection will determine its usefulness as a platform for evaluation, both in the short and long term future.

All topics created by participants will be used in the *Book Retrieval* and the *Focused Book Search* tasks. The final test collection will consist of only those topics for which relevance judgements will have been collected during the assessment phase. The same set of topics will be used for both retrieval tasks.

This document provides guidelines on how to create topics. Please read this document carefully (even if you participated in the track before as **there are significant changes from previous years**).

The main difference from last year is that participants are encouraged to use **Wikipedia** at different stages when preparing topics. The intuition behind the introduction of Wikipedia is twofold. First, Wikipedia articles often contain a reading list of books relevant to the general topic of the article, while they also often cite related books relevant to a specific statement in the article. This gives us an opportunity to create topics with real application. Second, we expect that browsing through Wikipedia entries could provide participants with suggestions about topics and their specific aspects of interest, and at the same time provide them with insights and relevant terminology to be used for better searches and refinements that should lead to a better mapping between topics and collection. However, please keep in mind the nature of the book collection. As the books in the corpus are out-of-copyright, and almost all books were written before 1930, Wikipedia articles on recent events or topics should be avoided.

Both the *Book Retrieval* and the *Focused Book Search* tasks are based on the assumption that searchers are seeking information on a given subject for a given task at hand. In the *Book Retrieval* task, the goal is to build a reading list on a given topic, where books that are focused on the topic – and cover all or most aspects of the topic – are sought. In the case of the Wikipedia user task, these books may then be listed, for example, in the references section of the Wikipedia page. In the *Focused Book Search* task, the goal is to find relevant information in books for a given topic or aspect of a topic. In the context of Wikipedia, these book parts may be cited, for example, in support of a given statement.

## 2 Requirements and deadline

Each participating organisation is asked to create a **minimum of 2 topics**, which should be submitted by **midnight on Wednesday, July 12, 2009** using the candidate topic submission form at:

<http://www.inex.otago.ac.nz/tracks/books/TopicDevelopment.asp>

The submission of additional topics is highly encouraged. Note that this will not impact on your assessment load later on, but will allow for greater flexibility during the assessment collection stage.

## 3 Topic definition

A topic reflects an *information need* that has arisen from a particular *user task*.

A user task may be to write an essay or an article on a topic. **A special user task this year is to find books or book parts that are relevant to segments of Wikipedia articles.** The segment may be a sentence, a paragraph, or even the whole article.

An information need may be generic or specific. Reflecting this, a topic may be of varying complexity and may comprise one or multiple aspects or sub-topics. We encourage participants to create multiple aspects of topics, where aspects should be focused (narrow). These focused aspects should have a limited number of relevant book parts (e.g., pages). This is to ease the relevance assessments later on and to ensure that judgements can be assumed to be more or less complete.

As an example, consider the task of writing an essay on Attila the Hun, or compiling a Wikipedia page on Leonardo da Vinci's inventions. Each of these topics includes a number of sub-topics, covering various aspects of the topic. For example, sub-topics of "Attila the Hun" include "Attila's marriages" and the "Hun invasion of Italy under Attila", etc. Aspects of "Leonardo da Vinci's inventions" include "musical instruments" and "hang glider", etc.

## 4 Topic Format

Topics are made up of several parts, each of which describe the same information need, but for different purposes and at different levels of detail. The Wikipedia fields, marked in red ink below, only need to be filled in if participants chose the Wikipedia user task or if they referred to Wikipedia while defining their topic.

<b>Task</b>	A detailed description of the task at hand for which information is sought, specifying the context, background, and the motivation for the need.
<b>Topic title</b>	The topic title serves as a summary of the user's information need. It will be used (in the automatic runs) as the query to search the book corpus.
<b>Topic Description</b>	A natural language description of the information need.
<b>Topic Narrative</b>	A detailed explanation of what information is sought and what is considered relevant or irrelevant.
<b>Wikipedia title</b>	The title of the Wikipedia page that relates to the topic. This may be the same as the topic title.
<b>Wikipedia URL</b>	The URL of the Wikipedia page that relates to the topic.
<b>Wikipedia text</b>	A copy of the textual content of the segment of the Wikipedia page on which the topic is based. This information may be used during in both the <i>Book Retrieval</i> and <i>Focused Book Search</i> tasks as context.

A topic may have any number of **aspects**, but must have a **minimum two**. Each aspect of a topic must be described in the following format:

<b>Aspect title</b>	The sub-topic title serves as a summary of the specific aspect of the user’s information need. It will be used (in the automatic runs) as the query to search the book corpus.
<b>Aspect narrative</b>	A detailed explanation of what information is sought and what is considered relevant or irrelevant.
<b>Wikipedia text</b>	A copy of the segment of the Wikipedia page that relates to this sub-topic.

Both the topic and aspect narratives must be clear and precise descriptions of the information need in order to unambiguously determine whether or not a given text fragment in a book fulfils the need. The narrative will be taken as the only true and accurate interpretation of the user’s needs. Relevance assessments will be made on compliance to the narrative alone.

An example topic is given below.

```

<topic>
<task> My task is to add citations to the Wikipedia article on Donations of Alexandria referring to books and parts of
books relevant to the topic and specific aspects of the topic. </task>
<title> Donations of Alexandria </title>
<description> I am looking for information on the political statement by Mark Antony that is referred to as the
Donations of Alexandria in which he distributed lands held by Rome and Parthia amongst Cleopatra
and their children. </description>
<narrative> Any information relating to the donations, their motivation and political background, and their
consequences are relevant. Data on what donations and titles were given to whom are relevant. Details
of the ceremony when the donations were announced are also relevant. Octavian’s response and the
political situation that arose as a result of the donations, leading to the civil war, are all relevant.
</narrative>
<wikipedia-title> Donations of Alexandria </wikipedia-title>
<wikipedia-url> http://en.wikipedia.org/wiki/Donations\_of\_Alexandria </wikipedia-url>
<wikipedia-text> [text of the Wikipedia page] </wikipedia-page>

<aspect>
<title> The Donations </title>
<narrative> Of relevance are the details of the donations, i.e., the ceremony and the particulars of who got
what. </narrative>
<wikipedia-text> [text of the “The Donations” section from the Wikipedia page] </wikipedia-text>
</aspect>

<aspect>
<title> Consequences </title>
<narrative> Any information that details how the donations were received and what reactions they generated, in
particular how Octavian reacted. Descriptions of the political atmosphere are relevant as well as events leading to the
civil war. The civil war itself is however not relevant. </narrative>
<wikipedia-text> [text of the “The Donations” section from the Wikipedia page] </wikipedia-text>
</aspect>
</topic>

```

## 5 Topic creation criteria

When creating topics, the following should be taken into consideration:

- Topics should reflect real information needs,
- Topics should have at least 2 but no more than 20 relevant books in the top 100 results,
- Topics should have at least 2 aspects,
- Each aspect of a topic should have at least 1 relevant book page, but no more than 100 pages,
- Topics and aspects should cover subjects for which the topic author will be able to provide relevance judgements.

## 6 Procedure for Topic Development

Please follow the steps below to create your topics and submit them using the Topic Submission Form on the INEX website (URL is given in Section 2, and the template is shown in Appendix B). ***It is crucial that all parts of a topic are carefully constructed!***

### Step 1: Initial Topic Statement and Task

Create a one or two sentence description of what information you are seeking and why, i.e., for the completion of what task you seek information. Write this in the *Initial Topic Statement* field. In the case of the Wikipedia user task, locate the Wikipedia page and the segment within it that is closest to your topic. Identify the different aspects (sub-topics) of your information need. In the case of the Wikipedia user task, for each aspect, identify the relevant segment of the Wikipedia page.

Describe your task, recording the context and the motivation for your information need. Write these in the *Task* field. Include a description of how would any found relevant information be used.

### Step 2: Collection Exploration

The goal of this phase is to obtain an estimate of the number of books and pages within books that are relevant to your topic and its aspects. This is also a crucial stage to evaluate whether the topic can be judged consistently.

We recommend that you start with the aspects and assess whether there are relevant book pages for these. You should find at least one relevant page for each aspect of your topic. However, you should revise a topic aspect, by making it more specific, if you find more than 100 relevant pages for it. Once you explored each aspect, you can derive an aggregate estimate of the number of relevant books for the whole topic. You should abandon a topic and start the process with a new topic if you find less than **2** or more than **20** relevant books in the top 100 results.

You can search the book collection using the book search engine available at [www.booksearch.org.uk](http://www.booksearch.org.uk), see Appendix C for a brief user description of the system. To search, enter your query terms into the search engine and browse the returned book results. You should try a number of different queries and mark any relevant book or page that you find. The system will record your judgements.

To assess the relevance of a book or a page, use the following working definition: Mark it relevant if it would be useful for achieving the goals of your task, e.g., if it contributes toward satisfying your information need. Note, however, that each result should be judged on its own merits. That is, information is still relevant even if it is the third time you have seen the same information. It is important that your judgment of relevance is consistent throughout this task.

When you finished with a topic, view your relevance assessment history report provided by the search engine (see Appendix C), and copy and paste your judgements into the Topic Submission Form. You may also want to save this information locally for future reference (or as backup). Before you start on a new topic, make sure you reset your relevance assessments log.

### Step 3: Narratives

By now you should have a clear idea of what information is available within the book corpus and what information you consider relevant or not for your chosen topic and its aspects. It is important that you record this knowledge in detail in the respective narrative fields of the online form. Record not only what information is being sought, but also what makes it relevant or irrelevant.

Make sure your description is exhaustive for two reasons: 1) this information will be **very important** for others to understand your information need; 2) you may not remember all of it months later when you need to provide relevance assessments.

### Step 4: Description and Titles

During the exploration phase your initial topic statement has likely evolved. Record in natural language the final version of your topic statement in the *Topic Description* field of the online form.

Compose the title for your topic and each aspect by recording the query words that you would use to search. Write these in the *Topic Title* and *Aspect Title* fields, respectively.

### **Step 5: Wikipedia Information**

If you are composing topics based on Wikipedia articles, then make sure you record the title and URL of the Wikipedia page and that you copy and paste the text of the relevant segment of the Wikipedia page and the specific sub-parts of the segment for the topic and each aspect of the topic, respectively.

### **Step 6: Finalize**

Go over the whole topic (including aspects) and make sure that it reflects exactly and concisely what your intentions were. It is important that all parts of the topic express the same information need, be it at different level of detail. Note that we aim to run a range of challenges using different parts of a topic, including the text of the Wikipedia page segments. For example, the whole segment may be used to search the book corpus.

### **Step 7: Topic Submission**

To submit your topic, fill out the online Topic Submission Form on the INEX website - see Section 2 for the URL.

After submitting a topic you will be asked to fill in a short questionnaire (should take no longer than 2-3 minutes). It is important that this is done as part of the topic submission as the questions relate to the individual topic just submitted. This is part of an effort to collect more context for INEX topics, thereby increasing the reusability of the test collection (see [1]).

## **7 Topic Selection**

There is no topic pruning or selection done by the organizers. This means that all topics that you create will be distributed to all participants and subsequently included in the test set to be used in the search phase. Thus, you are solely responsible for the quality of your own topics.

Please also note that details of the topic author will be included in the distributed topic information. This is to allow other participants to contact topic authors for clarification. This information, however, will not be included in the final test collection.

## **References**

[1] Kamps, J. and Larsen, B. (2006). Understanding Differences between Search Requests in XML Element Retrieval. In *Proceedings of the SIGIR 2006 Workshop on XML Element Retrieval Methodology*, p. 13-19. [<http://www.cs.otago.ac.nz/sigirmw/>]

## Appendix A: Sample base queries

Participants may choose from the following set of queries for which at least two relevant books exists in the corpus, and create topics around these. Please note, however, that the same rules and procedures of topic creation apply.

Alexandria	Immigrants	philosophers
Algebra	Immigration	Plant Genetics
American Colonies	Indian Myths	Prison Camps
American Presidents	Indian Summer	Psalms
Americus Vespucius	Indians	Pythagoras
Anabaptists	Israel	Quakers
Ancient History of the Earth	Jewish Theology	Queen Elizabeth
Atheism	Judaism	Ralph Waldo Emerson
Battle of Bull Run	Juggernaut	Rationalism
Buddha	Last Supper	Reformation
Calculus	Laws of Nature	Reminiscences
California	Life of Shakespeare	Republicanism
Catholic Church	Life of Ulysses S. Grant	Sabbath Hours
Catholicism	Logic	Saint Boniface
Charlemagne	Major religions of the world	Sermons
Christianity	Meditation	Sidney Lanier
Church of England	Metaphysical Philosophy	Sikh History
Church of Scotland	Metaphysics	Socratic method
Confederate States	Michelangelo	Spiritual Experience of St. Paul
Confucius	Middle Ages	Symphony
Crusades	Military	The Algebra of Logic
Descartes	Missionaries	The Fall of Mexico
Dreams	Mississippi River	The Sistine Chapel
Emancipation Proclamation	Moravian Church	The War in Cuba
Eucharist	Motherhood	The Works of Daniel Webster
European Settlements in America	Muhammad	Time of Nero
Geometry	Mysticism	Treaty of Washington
Great Rebellion	Myths Every Child Should Know	Vedanta Philosophy
Hebrew culture	Napoleon	Vedas
Hebrew History	New Hampshire	War for Independence
History of America	Niagara Falls	Wind and Weather
History of Mexico	North Pacific Coast	Wine Industry
History of the Early Church	Ohio River	Works of Aristotle
History of the Southwest	Old Testament	
Holy Scriptures	Paganism	
	Passover	

# Appendix B: Topic Submission Template

## Candidate Topic

### Step 1: Initial Topic Statement

### \* Step 2: Book Corpus Exploration Phase

Found relevant books and book pages. You can copy and paste the output of the Book Search Engine's Topic Creation Report here.

### \* Step 3: Finalised Topic

Task:

Topic title:

Description:

Narrative:

Wikipedia article title:

[Empty rectangular box]

Wikipedia article title:

[Empty rectangular box]

Wikipedia article text:

[Empty rectangular box]

**\* Aspects**

Title:

[Empty rectangular box]

Narrative:

[Empty rectangular box]

Wikipedia segment text:

[Empty rectangular box]

....

## Appendix C: Book Search System

The search system is provided to INEX participants with the aim to aid topic creation (and to collect relevance judgements at a later stage). It is available at: [www.booksearch.org.uk](http://www.booksearch.org.uk)

### Logging in

Start by logging in using your INEX login credentials by following the Login link in the top right corner of the start-up window. If you have problems logging in, first check that you are registered for the Book Track at <http://www.inex.otago.ac.nz/people/participants.asp>. If you are registered, then please email Gabriella Kazai at [gabkaz@microsoft.com](mailto:gabkaz@microsoft.com).

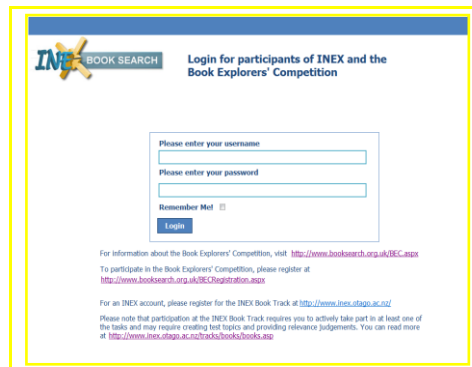


Figure 1. Login screen

### Main menu

Once you are logged in, you will be displayed a list of menu options. Please select the [Topic creation](#) link.

### Search box

Next, you will be displayed the main search screen, see Figure 2, with the search box at the top. The left hand side lists the Library of Congress categories extracted from MARC records associated with books in the corpus. You can navigate these categories to explore the keyword clouds associated with the different categories. Keyword clouds will appear in the middle part of the window, representing the most important terms from the subset of books that belong to a given category. You can click any of these keywords to add them to the search box.



Figure 2. Entry page for topic creation after login.

## Search results

Search results will be returned as a ranked list of books, see Figure 3. For each book, title, author and publication information is shown as well as a keyword cloud generated from the content of the book. Also shown are selected entries from the table of contents (if available) and a snippet from the highest ranking page of the book.

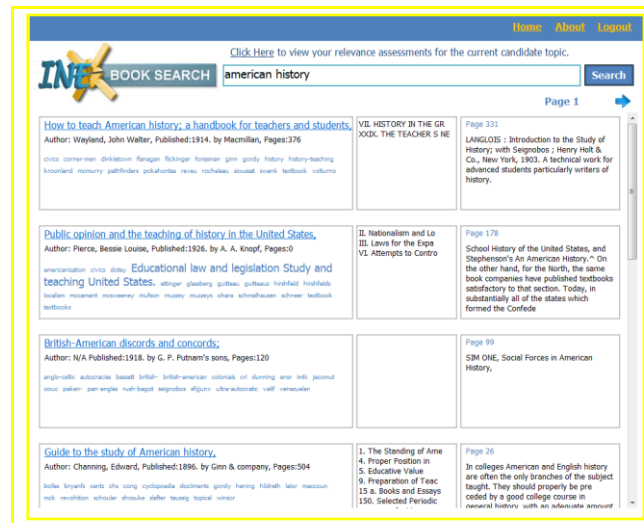


Figure 3. Search results.

## Book viewer window

To view a book, click its title. The book viewer window will appear, see Figures 4 and 5. The book metadata and the keyword clouds are repeated in the top left corner. You can click any of the keywords to add to the Search inside input box. A search will return a ranking of book pages, displayed in the Pages tab (see Figure 5). To view a page, click the Page number hyperlink.

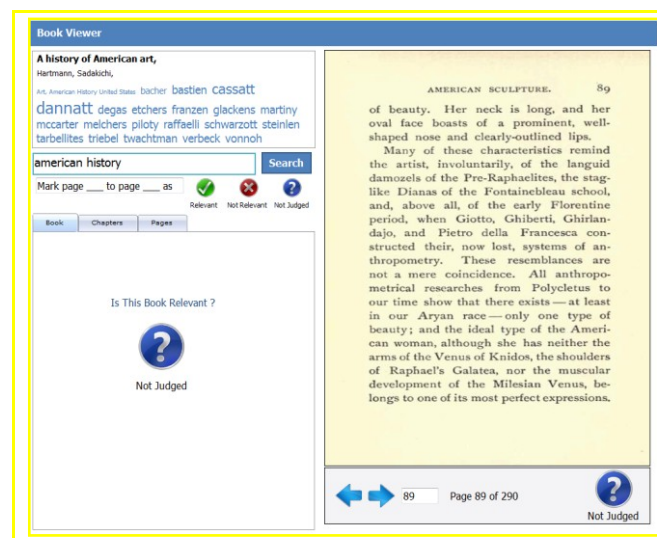


Figure 4. Book viewer window, showing the Book tab as active.

To navigate inside the book, you can use the previous/next arrows below the page image, or type in a page number and hit Return.

To judge a whole book relevant or irrelevant, click the question mark icon displayed in the Book tab. To judge the currently viewed page, click the question mark icon shown below the page image. Each relevance assessment icon rotates through the following settings: un-judged (default), relevant, and

irrelevant. You can also mark a range of pages as relevant by entering the physical page number<sup>1</sup> of the first and last page of the range using the format: e.g., 003-017.

You may assess any number of books and pages inside a book, and run a number of different queries. The system will record your query and the books/pages that you marked relevant.

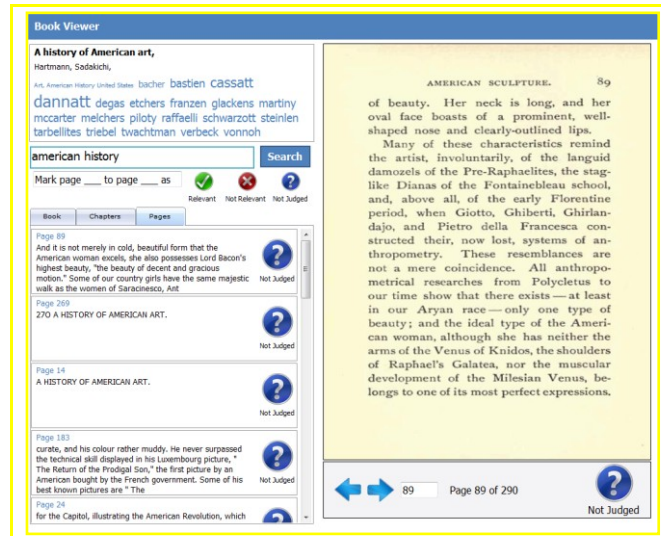


Figure 5. Book viewer window, showing the Pages tab as active.

## Relevance assessment session log

The assessment log shows your queries and the books/pages that you marked relevant, see Figure 6. You can access this log any time by clicking the hyperlink above the search box on the search results screen (see Figures 2 and 3). You may also keep this window open at all times (in a new window) and click refresh time to time to update its contents.

Please copy and paste this information into the Topic Submission Form (see Section 2 for the URL).

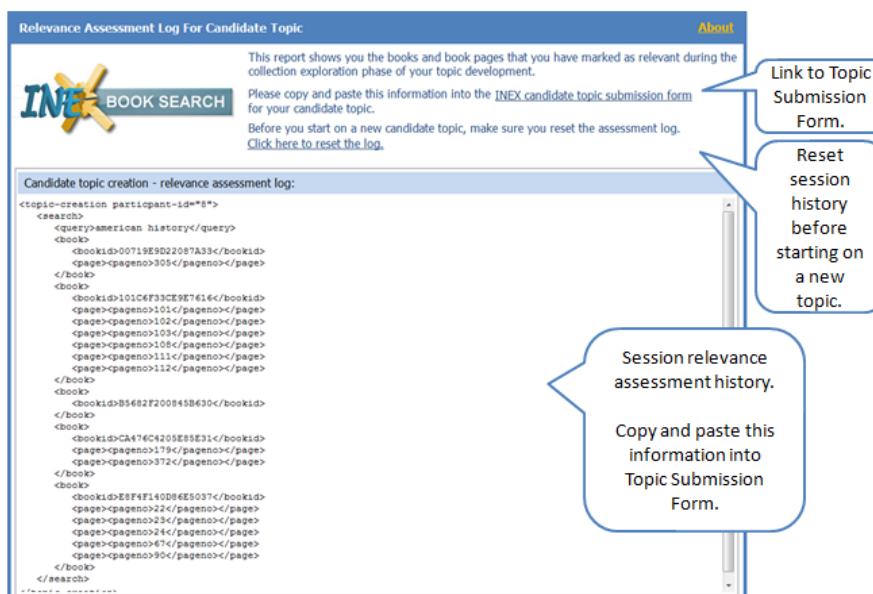


Figure 6. Assessment session log.

<sup>1</sup> Physical page numbers start at 1 on the first page of the book. They are different from the logical page numbers that are printed on a page image!