Best Entry Points for Structured Document Retrieval – Part I: Characteristics

Jane Reid[a], Mounia Lalmas[a], Karen Finesilver[b] and **Morten Hertzum[c]**

[a]Computer Science, Queen Mary, University of London, E1 4NS, UK
{jane,mounia}@dcs.qmul.ac.uk

[b]Open & Distance Learning Unit,Queen MaryUniversity of London, E1 4NS, UK
karen@odl.qmul.ac.uk

[c]Computer Science, Roskilde University, DK-4000, Denmark
mhz@ruc.dk

Abstract
Structured document retrieval makes use of document components as the basis of the retrieval process, rather than complete documents. The inherent relationships between these components make it vital to support users' natural browsing behaviour in order to offer effective and efficient access to structured documents. This paper examines the concept of best entry points, which are document components from which the user can browse to obtain optimal access to relevant document components. In particular this paper investigates the basic characteristics of best entry points.

Keywords: Structured document retrieval, focussed retrieval, best entry points, user studies.

1. Introduction
Document collections often display hierarchical structural characteristics. For example, a report may contain sections and subsections, each of which is composed of paragraphs. Structured document retrieval (SDR) attempts to exploit such structural information by retrieving documents based on combined structure and content information e.g. (Brin & Page, 1998; Frisse, 1988; Kotsakis, 2002; Myaneg et al, 1998; Wilkinson, 1994)

Structural information can be exploited at several stages of the information retrieval process. Firstly, it can be used at the indexing stage, where document components are identified and indexed as separate, but related, units (Cleveland, Cleveland & Wise, 1984; Tenopir & Ro, 1990). Secondly, structural information can be used at the retrieval stage by allowing document components of varying granularity to be retrieved, thus cutting down the amount of effort a user spends in pinpointing relevant information e.g. (Fuhr & Gro□johann, 2001; Rölleke, 1999). The issue of retrieval granularity has been addressed by many of the approaches developed for retrieval of documents in XML format; see (Fuhr, Malik & Lalmas, 2004). Thirdly, such information can be used at the results presentation stage by making structural relationships explicit to the user, thus reducing time and disorientation caused by lack of proximity of related document components in results interfaces, e.g. fisheye views to enable effective browsing of large documents (Furnas, 1999); expand-collapse functionality to support focus on, and movement between, particular structural elements of documents e.g. (Hertzum & Frøkjær, 1996); and use of clustering or sub-lists of related objects by web search engines (e.g. Google). Results presentation may also be focussed by presenting only selected document components in the interface, rather than all relevant document components: this approach, known as ***focused retrieval***, is the topic of this paper.

Focussed retrieval acknowledges the importance of users' natural browsing behaviour, and combines the browsing and querying paradigms to return ***best entry points*** to structured documents. There have been several interpretations of the broad concept of best entry point (BEP) in different fields. In web retrieval, in particular in the topic distillation task of the TREC web track[1] a ***good entry page*** to a web site is a page that is principally devoted to a topic and is not part of a larger site also principally devoted to the topic (Craswell et al, 2003). In INEX, the evaluation initiative for XML retrieval[2], a document component that is both ***exhaustive*** (describing the extent to which an XML element covers or discusses the topic of request) and ***specific*** (describing the extent to which an XML element is focused on the topic of request) to the information need could

[1] http://es.csiro.au/TRECWeb/guidelines_2003.html
[2] http://www.is.informatik.uni-duisburg.de/projects/inex/index.html.en

be considered an equivalent concept (Fuhr, Malik & Lalmas, 2004). A third interpretation, which is the one followed in this work, is that of (Chiaramella, Mulhem & Fourel, 1996) in the context of hypermedia multimedia retrieval. Here, a BEP is a document component from which the user can obtain optimal access, by browsing, to relevant document components. In all these interpretations, the use of BEPs in place of relevant document components as the basic units of the results list is intended to support the information searching behaviour of users, and enable them to gain more effective and efficient access to relevant information items.

In (Kazai, Lalmas & Reid, 2003) a test collection for the focussed retrieval of structured documents was constructed. This test collection consisted of a structured document collection, a set of queries adapted to the SDR task, and sets of corresponding relevance assessments and BEPs elicited from experimental participants. Initial analysis of the participants' BEP selection criteria led to the development of BEP retrieval algorithms (Kazai, Lalmas & Rölleke, 2001). These algorithms were implemented and empirically evaluated (Kazai, Lalmas & Rölleke, 2002), but were found to yield poor performance. We then carried out two further small-scale studies in order to explore, in more detail, particular aspects of the issue of automatic BEP identification (Lalmas & Reid, 2003). The first study employed algorithms incorporating different combinations of information related to the individual components themselves, their structural level, and their hierarchically related components, in order to identify BEPs automatically from a ranked results list of document components (Pithia, 2002). The second study employed rules derived from statistical analysis of the experimental data in order to identify BEPs automatically from known relevant document components (Srigananathan, 2002). Overall, the results of both these studies were poor, but it was concluded that combination of information related to, primarily, the hierarchical relationships between components and, secondarily, structural level, may provide the most promising method of automatic BEP identification.

In summary, the findings from these studies highlighted the need for further investigation of the **characteristics** of BEPs in order to inform the development of more sophisticated algorithms for the automatic identification of BEPs. In this paper, we describe a study that was designed to explore the characteristics of BEPs, in particular in comparison to relevant components. Results analysis is also performed across different query categories (factual/essay-topic; content-only/content-and-structure).

The remainder of the paper is organised as follows. In Section 2, we describe the Shakespeare test collection data, upon which our study is based. Sections 3 and 4 examine two specific BEP characteristics: the relationship between BEPs and relevant objects, and agreement between BEPs and relevant objects. In both these analyses, we discuss the effect of three factors: query complexity (factual/essay-topic queries), query type (content-only/content-and-structure queries), and structural level. In Section 5, we report preliminary findings from a further small-scale user study designed to identify and analyse BEPs for the INEX data set (Kazai et al, 2004). We conclude the paper with a summary of our findings and proposals for future work in Section 6.

2. Shakespeare user study data

In this section, we introduce the basic elements of the Shakespeare user study data: the document collection, queries, relevance assessments and BEPs[3]. Eleven students from the undergraduate BA in English Literature and Drama at Queen Mary, University of London participated in the construction of the test collection (i.e. providing queries, relevance assessments and BEP assessments). The experimental methodology used to acquire the data is fully described in (Kazai, Lalmas & Reid, 2003).

2.1. Document collection

The document collection consists of 12 Shakespeare plays marked up in XML by Jon Bosak, and available from the web at http://www.ibiblio.org/bosak/. The 12 plays are Antony and Cleopatra, A Midsummer Night's Dream, Hamlet, Julius Caesar, King Lear, Macbeth, Much Ado About Nothing, Othello, Romeo and Juliet, The Tempest, Troilus and Cressida, and Twelfth Night. Each piece of content enclosed by XML tags is a retrievable element. The maximum depth of nested XML elements is 6. Each play contains, on average, 5 096 elements, 5 acts, 21 scenes, 892 speeches and 3 311 lines.

2.2. Queries

The test collection includes 43 queries. Each query relates to a single play in the collection (average 3.58 queries per play), and was produced by an experimental participant as the result of a real information need. The queries are of varying complexity:

[3] The data are available for public use, in the form of a test collection, at http://qmir.dcs.qmul.ac.uk/Focus

- **Factual.** Generally, a small number of short, simple elements provide the answer to such queries. An example query is "How old is Juliet?". There is a maximum of one factual query per play.
- **Essay-topic.** Generally, reference will have to be made to many, complex elements to find the answer to such queries. An example query is "Describe the character of Lady Macbeth".

The queries also reflect the additional functionality of structured query languages (e.g. XPath[3], XQuery[4]), i.e. that it is possible to query by structure. Each query, therefore, belongs to one of two types:

- **Content-only.** This is the standard type of query in information retrieval: such queries describe a topic of interest to the user.
- **Content-and-structure.** This type of query combines topic and structural requirements. An example of such a query is "Retrieve the title and first speech of scenes where the portrayal of the Trojans is compared with that of the Greeks".

The distribution of queries across query categories is shown in Table 1.

Table 1. Distribution of queries across query categories.

	Content-only	Content-and-structure	Total
Factual	9	2	11
Essay-topic	26	6	32
Total	35	8	43

2.3. Relevance assessments

In the Shakespeare study, the relevance assessments are binary, and indicate which elements the experimental participants considered they would consult (read or reference) in order to answer the given query. All relevance assessments were made at the lowest structural level, referred to as leaf level elements. This structural level corresponds mostly to the Line level of the collection structure. Participants were asked to highlight the lines that they considered as relevant to their information need, as expressed in the corresponding queries. This was considered a feasible approach, since the test collection is reasonably small.

Each query was judged by 2-4 participants (average 2.72 judges per query), resulting in 117 relevance assessment sets, containing a total of 6 296 leaf level XML elements, for the 43 queries (average 53.81 relevant leaf level elements per set and 146.42 relevant leaf level elements per query). When these multiple sets of relevance assessments are pooled for each query, and duplicate elements removed, this gives a total of 4 894 unique leaf level XML elements in 43 query sets (average 113.81 elements per query). Since there is only one relevant play for each query, and given that a play contains, on average, 5 096 elements, the relevant elements for a given query represent 2.23% of the play.

2.4. Best entry points (BEPs)

In our work, BEPs are document components from which the user can obtain optimal access, by browsing, to relevant document components. In this case, participants were asked to identify as BEPs those elements that they would prefer to be retrieved in response to the given query. This was done through the use of an interface, which allowed participants to view the relevant document components and navigate through the collection. It should be noted that BEPs can belong to any structural level, and are not always elements that have been judged relevant, but are, in some cases, non-relevant container or contained elements.

The participants identified a total of 928 BEPs for the 43 queries, in 117 sets (average 7.93 BEPs per set and 21.58 BEPs per query). This number is reduced to 552 by removal of duplicate elements (average 12.84 BEPs per query). The BEPs for each query, as judged by each participant, can then be merged to form the final set of BEPs; usually only elements judged as BEPs by the majority of the participants would be included. This is to avoid the problem of multiple BEPs representing the same cluster of relevant elements, e.g. two individual participants choosing two different lines of the same speech as BEPs. The total number of BEPs in this reduced set is 521 (average 12.12 BEPs per query).

[3] http://www.w3.org/TR/xpath
[4] http://www.w3.org/XML/Query

3. Relationship between BEPs and relevant objects (ROs)

As stated in section 2.4, BEPs can belong to any structural level, while relevance assessments are all leaf-level. In Table 2, we show the distribution of BEPs across structural levels. For this analysis, the (non-merged) set of unique BEPs (552 BEPs in total) was used, and Line and StageDir objects were collapsed into the leaf object type. The raw values and percentages for each category were calculated by averaging across all queries belonging to that category. It should be noted that there are no BEPs at act level, and only one at play level.

Table 2. Distribution of BEPs for different query categories across structural levels.

Query category	Leaf	Speech	Scene	Act	Play	Other	Total
Factual	4.36 (63%)	3.18 (25%)	0.27 (2%)	0 (0%)	0 (0%)	0.45 (10%)	8.27 (100%)
Essay-topic	6.00 (38%)	7.38 (54%)	0.97 (8%)	0 (0%)	0.03 (0%)	0.03 (0%)	14.41 (100%)
Content-only	6.26 (47%)	6.66 (44%)	0.80 (6%)	0 (0%)	0 (0%)	0.06 (3%)	13.77 (100%)
Content-and-structure	2.63 (34%)	4.75 (57%)	0.75 (5%)	0 (0%)	0.13 (1%)	0.50 (3%)	8.75 (100%)
Overall average	5.58 (44%)	6.30 (49%)	0.79 (6%)	0 (0%)	0.02 (0%)	0.14 (1%)	12.84 (100%)

The total number of BEPs for each query category differs considerably, with essay-type and content-only queries having more BEPs than factual and content-and-structure queries. However, the relative distribution of the BEPs is rather similar across query types and complexities. Overall, the majority of the BEPs were selected from the leaf or speech levels, together accounting for 93% of all BEPs. This suggests that participants generally show a preference for smaller contexts. Examining the effect of query complexity, factual queries have more BEPs at leaf level and fewer at speech and scene levels. Examining the effect of query type, content-and-structure queries have more BEPs at speech level and less at leaf level.

Table 3 shows the distribution of ROs across the two main leaf levels (Line and StageDir) and for other types of objects (Other), e.g. Persona. The majority of ROs for all query categories are of Line type, as would be expected. Most of the query categories also have a similar total number of ROs per query, with the exception of factual queries, which have fewer. Examining the effect of query complexity, factual queries show a comparatively high percentage of Other objects; however, this result was heavily influenced by one individual query (query 34, relating to Macbeth: "Who does the ghost appear to and does he/she/they ever see the ghost again?). This query could only be answered by reference to Speaker objects (part of the Other object type). Examining the effect of query type, content-and-structure queries show a very high percentage of Line objects.

Table 3. Distribution of ROs for different query categories across structural levels.

Query category	Line	Stagedir	Other	Total
Factual	75.73 (93%)	1.73 (2%)	2.00 (5%)	79.45 (100%)
Essay-topic	122.69 (98%)	2.75(2%)	0.19 (0%)	125.63 (100%)
Content-only	112.03 (96%)	2.94 (2%)	0.60 (2%)	115.57 (100%)
Content-and-structure	104.75 (99%)	0.50 (1%)	0.88 (0%)	106.13 (100%)
Overall average	110.67 (97%)	2.49 (2%)	0.65 (1%)	113.81 (100%)

The above results for BEPs and ROs can be explained by the particular properties of factual and content-and-structure queries. Factual queries involve a question-answering process, rather than an evidence-gathering one. The "answer" to factual queries is, therefore, likely to be contained in fewer, lower-level contexts, thus explaining the lower total number of ROs, and the higher proportion of ROs at leaf level rather than speech level. Content-and-structure queries make explicit reference to a structural requirement, which imposes an additional constraint on the selection process, i.e. that ROs must appear at a particular structural level, often the speech level in this collection. This indicates that there will be a lower number of candidate BEPs and ROs, and a higher proportion of BEPs at speech level rather than leaf level. In addition, 6 of the 8 content-and-structure queries were also essay-topic queries (evidence-gathering, rather than fact-finding), so, again, are more likely to involve ROs at a structural level above leaf level.

We can observe an interesting pattern in the identification of BEPs from ROs. BEPs tended to be chosen at the same structural level as the corresponding ROs, or one level above in the hierarchy. Participants often specified a sequence of consecutive lines as being relevant, and then chose the first line of this sequence as being the BEP: this was the case in 87% of such linear chains. In other cases, the level above was generally chosen as the BEP level, most often involving a straightforward move from sub-speech sequence to entire speech.

In the Shakespeare study, this result may have been influenced by the fact that experimental participants were instructed to specify ROs at leaf level. However, these results can also be compared with data on the relationship of relevant items to preferred entry points gathered during the Tess study (Hertzum & Frøkjær, 1996; Hertzum, Lalmas, & Frøkjaer; 2001). In the Tess study, 83 participants solved a number of information retrieval tasks in relation to the development of graphical user interfaces in the X Window System, a domain very different from Shakespeare's plays. The documentation necessary to solve the tasks was three manuals comprising 774 pages of text. In 51% of tasks, the Tess participants identified the answer to a task by searching and, therefore, preferred an entry point into the documentation that was equal to a relevant item. However, in as much as 21% of the tasks the participants chose to enter the documentation at a higher level in the text hierarchy than the text containing the answer[5]. In these cases the entry point preferred by the participants was at a structural level above the relevant items, which were then identified by reading rather than searching. Access to the documentation was most frequently at the 2nd and 3rd levels (subchapters and sub subchapters), which is an average level slightly above that of the ROs.

The results from these two studies agree, therefore, that participants generally display a preference for low-level contexts, and usually at the same level as that of the ROs, or the level immediately above. This implies that BEPs should not simply correspond to the elements containing the answers to the tasks, but rather must strike an appropriate balance between searching and reading.

Finally, we analysed the relationship of BEPs to ROs. Table 4 shows average figures for ROs and BEPs for all query categories, and the degree of reduction for each category. The reduction for an individual query is calculated as (1 – number of BEPs / number of ROs), and the reduction for a query category is calculated as the average of the reduction values for all queries belonging to that category.

Table 4. Relationship of BEPs to ROs for different query categories.

Query category	Average number of ROs	Average number of BEPs	Average reduction
Factual	79.45	8.27	52%
Essay-topic	125.63	14.41	87%
Content-only	115.57	13.77	77%
Content-and-structure	106.13	8.75	82%
Overall average	113.81	12.84	78%

The number of BEPs is, for all query categories, less than the number of ROs. The reduction is generally large, with 30 of the 43 queries having a reduction of more than 80%. The reduction values are similar across all query categories, apart from factual queries, which show a considerably lower reduction percentage. This could be expected, since factual queries generally have fewer ROs (79.45 for factual queries compared with 113.81 across all categories), often in the form of individual lines well distributed throughout the text; the reduction process is thus likely to eliminate fewer objects.

Analysis was then performed on the reduction process at different structural levels. Although ROs were all leaf level (mostly Line objects; see Table 3), *extrapolated ROs* were created by assuming relevance at the structural level above that of the original RO (Table 5). An optimistic relevance extrapolation strategy was used (Rölleke et al, 2002); for example, if one line was marked as relevant, relevance was extrapolated to the (complete) speech containing that line, and so on[6]. Since BEPs could belong to any structural level, no extrapolation was performed for them. Table 6 shows the reduction levels from (leaf and extrapolated) ROs to BEPs across structural levels.

The small number of ROs at speech level is caused by a high number of leaf level ROs occurring in the same container object (speech) as other leaf level ROs for the same query. For example, if one participant chooses the first line of an individual speech as an RO, and another participant chooses the second and third lines of the same speech, all these would be extrapolated to the same RO at speech level.

[5] For the remaining 28% of the tasks the Tess subjects failed to identify the relevant parts of the documentation.

[6] A document component being relevant implies that its parent element is also relevant to some extent. Although not explicitly stated during this particular study, this view was later adopted in the context of INEX.

Table 5. Leaf and extrapolated ROs for different query categories across structural levels.

Query category	Leaf	Speech	Scene	Act	Play
Factual	79.45	19.00	3.08	1.83	1
Essay-topic	125.63	34.26	5.03	2.84	1
Content-only	115.57	32.11	4.80	2.57	1
Content-and-structure	106.13	20.75	3.13	2.50	1
Overall average	113.81	30.00	4.49	2.56	1

Table 6. Reduction of leaf and extrapolated ROs to BEPs across structural levels.

Query category	Leaf	Speech	Scene	Act	Play
Factual	61%	76%	95%	100%	100%
Essay-topic	94%	71%	81%	100%	97%
Content-only	85%	72%	85%	100%	100%
Content-and-structure	88%	73%	79%	100%	88%
Overall average	86%	73%	84%	100%	98%

The reduction can be seen to be high across all query categories and structural levels. In fact, there was only one instance (query 2 at scene level) where the reduction was negative, i.e. there were more BEPs than ROs. This was due to a large number of relevant speeches occurring in the same scenes, leading to the choice of container scenes as BEPs, rather than the individual speeches. It is noticeable that the reduction is smaller for factual queries at leaf level; this could be expected, due to the smaller number of ROs for this category, and the high proportion of single-line objects considered relevant. We can see from these results that the choice of BEPs is not simply a function of the relevance of the contained objects: this analysis indicates a more complex view of relevance in the context of structured documents.

4. Agreement for BEPs and ROs

The second factor that we investigated in this study is agreement between participants in their choice of both BEPs and (leaf and extrapolated) ROs. Agreement was calculated for each query in terms of overlap, i.e. the size of the intersection of the BEP (or RO) sets for that query divided by the size of the union of the BEP (or corresponding RO) sets for that query (Voorhees, 1998). Unanimous agreement values were also calculated, i.e. where a BEP (or RO) must be judged relevant by all judges to be included in the intersection. The agreement for a query category was calculated as the average of the agreement levels for all queries belonging to that category. Results for BEP agreement can be found in Tables 7 (general) and 8 (unanimous), and for RO agreement in Tables 9 (general) and 10 (unanimous). In this analysis, leaf level includes all Line, Stagedir and Other objects. BEP agreement was also calculated across *all* BEPs (this was not done for ROs, since all assessments were made at leaf level, and relevance simply extrapolated to other levels).

BEP agreement (both general and unanimous) generally increases from leaf to speech level and then falls off at higher structural levels. The exception to this pattern is factual queries, which show a higher level of agreement at leaf level than speech level, before rising again at scene level.

RO agreement (both general and unanimous) increases consistently with structural level. This implies that participants may not always agree on the exact context of the RO, but tend to agree more on the general area in which the ROs can be found. The results show that query type and complexity do not have a strong effect on RO agreement, although factual queries show a slight increase in RO agreement at most structural levels. Unanimous agreement is higher for factual and content-and-structure queries, especially at lower structural levels. This result may be due, again, to the additional constraints these query categories impose, which leave less freedom for individual interpretation.

Comparing BEP and RO agreement, BEP agreement is better for all categories at leaf and speech level. For other levels, the opposite trend can generally be observed. This result may, however, be heavily influenced by the reduced number of BEPs at these higher levels. Another, related reason for the result may be the optimistic method of relevance extrapolation used to calculate RO agreement. This implies that there will be more ROs at higher structural levels, and the number of BEPs at higher structural levels may thus appear artificially low in comparison.

Table 7. General BEP agreement for different query categories across structural levels.

Query category	Leaf	Speech	Scene	Act	Play	All BEPs
Factual	63%	52%	67%	-	-	67%
Essay-topic	46%	62%	41%	-	0%	57%
Content-only	55%	60%	45%	-	-	72%
Content-and-structure	35%	59%	50%	-	0%	53%
Overall average	49%	58%	51%	-	0%	60%

Table 8. Unanimous BEP agreement for different query categories across structural levels.

Query category	Leaf	Speech	Scene	Act	Play	All BEPs
Factual	54%	11%	33%	-	-	52%
Essay-topic	30%	47%	22%	-	0%	47%
Content-only	40%	39%	22%	-	-	48%
Content-and-structure	26%	45%	33%	-	0%	48%
Overall average	37%	40%	24%	-	0%	48%

Table 9. General (leaf and extrapolated) RO agreement for different query categories across structural levels.

Query category	Leaf	Speech	Scene	Act	Play
Factual	35%	43%	59%	84%	100%
Essay-topic	27%	30%	68%	76%	100%
Content-only	29%	35%	65%	80%	100%
Content-and-structure	30%	30%	63%	73%	100%
Overall average	31%	35%	64%	78%	100%

Table 10. Unanimous (leaf and extrapolated) RO agreement for different query categories across structural levels.

Query category	Leaf	Speech	Scene	Act	Play
Factual	28%	34%	51%	76%	100%
Essay-topic	11%	14%	51%	66%	100%
Content-only	14%	18%	51%	71%	100%
Content-and-structure	24%	25%	52%	60%	100%
Overall average	16%	19%	51%	69%	100%

Table 11. Proportion of BEPs selected by individual participants across structural level.

Participant	Leaf	Speech	Scene	Act	Play	Other	All BEPs
1	60%	53%	70%	-	0%	100%	68%
2	54%	67%	72%	-	0%	67%	68%
3	28%	50%	48%	-	100%	-	44%
4	78%	61%	19%	-	0%	100%	72%
5	54%	69%	92%	-	-	67%	69%
6	62%	77%	67%	-	-	83%	75%
7	77%	68%	37%	-	-	-	75%
8	44%	65%	75%	-	-	100%	62%
9	71%	73%	50%	-	-	-	74%
10	75%	65%	25%	-	-	-	75%
11	49%	40%	96%	-	-	-	56%
Overall average	59%	63%	59%	-	25%	86%	67%

It should be noted that such work relies heavily on the quality of the data obtained from the experimental participants. In the Shakespeare user study, participants were aided in their selection of BEPs by the use of a user interface, which explicitly showed both the structure and content of the plays, and clearly highlighted the elements that had been marked relevant by at

least one participant (Kazai, Lalmas & Reid, 2003). Despite this, and the comparatively high level of resulting BEP agreement (Tables 7 and 8), there is still considerable variation between individual participants in terms of the proportion of (unique, non-merged) BEPs they selected for the queries they performed (Table 11). Although the average percentage (across all participants) is similar, especially at the lower structural levels, there is considerable variation among participants in terms of the overall percentage of BEPs chosen and bias towards structural level.

5. BEPS in INEX

INEX, the INitiative for the Evaluation of XML Retrieval (Kazai et al, 2004) provides an infrastructure (in the form of a large XML test collection) and methodology (e.g. appropriate scoring methods) for the evaluation of content-oriented retrieval of XML documents. The test collection contains a document collection, together with sets of topics and corresponding relevance assessments produced by the organisations participating in the initiative.

The INEX document collection is composed of full-text articles marked up in XML: the collection currently includes 12 107 articles of varying length from the IEEE Computer Society's publications (12 magazines and 6 transactions), covering the period 1995-2002. An article contains, on average, 1 532 XML elements, and the average depth of nested XML elements is 6.9. Typically, the overall structure of an article consists of a front matter, a body and a back matter. The front matter contains the article's metadata, e.g. title, author, publication information, and abstract. The body is structured in sections, sub-sections, and sub-sub-sections. These logical units start with a title, and contain a number of paragraphs, tables, figures, item lists, in-text references (citations), etc. The back matter includes a bibliography and further information about the article's authors.

Based on the subject domain of this document collection, each participating organisation created a set of candidate content-only and content-and-structure topics: these topics were designed to reflect the information needs of real users and the type of service that an operational system might provide. The topic format and development procedure were based on TREC guidelines, which were modified to allow for the specification of structural conditions (e.g. what type of components to return to the user) in content-and-structure queries.

In INEX 2002, a final set of 60 topics (composed of 30 content-only and 30 content-and-structure) was selected from the set of candidate topics, and distributed to the participating organisations. For each topic, participants generated queries from any part of the topic structure, except the narrative, and used these to retrieve ranked lists of XML elements from the document collection. The participants' submissions were then merged to form the pool of elements for assessment. During the assessment process, each topic and its pool of elements were assigned to a participating organisation. Where possible, the topic author was asked to perform the assessment; remaining topics were assigned, on a voluntary basis, to a participating organisation with expertise in the relevant subject area.

The assessment process was performed using *two* dimensions: a four-point scale of topical relevance, which describes the extent to which the information contained in a document component is relevant to the topic, and a four-category grouping for component coverage, which describes how much of the document component is relevant to the topic[7]. All ascendant components (parents and upwards) and descendant components (children and downwards) were also judged until irrelevant components and/or components with no coverage were encountered. Thus, the set of relevance assessments for an INEX 2002 topic includes all document components that have been judged not to be irrelevant (on the topical relevance dimension) and not to have no coverage (on the document coverage dimension). Assessments were recorded using an on-line assessment system, which allowed users to view the pooled result set of a given topic (listed in alphabetical order), browse the document collection, and view articles and result elements both in XML format (i.e. showing the tags) and document format (i.e. formatted for ease of reading). Other features of the on-line system included keyword highlighting and consistency checking of the assessments. Assessments were provided for 54 of the 60 topics, covering a total of 48 849 articles.

The organisations participating in INEX 2002 were not asked to identify BEPs. However, Ghuman (2004) subsequently carried out a small-scale experimental study using the INEX 2002 test collection, with the aim of making some preliminary observations about the concept of BEPs, as interpreted in this paper (i.e. entry points in the articles from which users could obtain optimal access to relevant XML elements). This study involved four undergraduate Computer Science students from Queen Mary, University of London, who were asked to identify BEPs for a single content-only topic, namely topic 47

[7] INEX 2003 and INEX 2004 employ a modified definition of relevance, but still based on two dimensions: exhaustivity and specificity (as described in Section 1). See (Fuhr, Malik & Lalmas, 2004) for full details.

(Figure 3). Topic 47 was chosen because it had a manageable number of relevance assessments (relevant objects or ROs) for the size of the study, which was carried out as part of a final year undergraduate student project. Further analysis, focussing on issues also examined in the context of the Shakespeare data, was then carried out. Although it is clearly not possible to generalise from these limited results, the analysis does highlight interesting areas for future exploration.

```
- <INEX-Topic topic-id="47" query-type="CO" ct-no="078">
   - <Title>
       <cw> concurrency control semantic transaction management
            application performance benefit </cw>
    </Title>
    <Description> What are the benefits achieved by deploying semantic
        transaction management techniques. </Description>
    <Narrative> Relevant documents/components are those that report on
        performance improvements with information systems - especially
        database systems - when using semantic transaction management
        as opposed to conventional transaction management such as two-
        phase locking. The documents/components should have an
        analytical investigation, a simulation or performance results
        from a prototype system. </Narrative>
    <Keywords> "concurrency control" "semantic transaction
        management" "application" "performance benefit" "prototype"
        "simulation" "analysis" </Keywords>
 </INEX-Topic>
```

Figure 3. Topic 47 from the INEX 2002 test collection, used in (Ghuman, 2004).

The examination of BEP characteristics and agreement was carried out according to structural level, where the level reflects the position of the given document component in the hierarchy of the article in which it appears, as defined by the INEX collection DTD. As described above, an article is generally divided into front matter (e.g. keywords, abstract), body (consisting of sections, sub-sections, paragraphs, etc) and back matter (e.g. bibliography). Thus, for example, a complete article would correspond to level 1, the body of that article to level 2, an individual section within the body to level 3, a paragraph of that section to level 4, and a sentence within that paragraph to level 5. None of the ROs or BEPs for the chosen topic were at a level lower than 5 in the XML structure.

Table 12. Relationship of INEX ROs to BEPs across structural levels.

	Level 1	Level 2	Level 3	Level 4	Level 5	All levels
Number of ROs	59 (16%)	29 (8%)	241 (67%)	28 (8%)	4 (1%)	361 (100%)
Number of BEPs	16 (7%)	3 (1%)	137 (57%)	81 (34%)	2 (1%)	239 (100%)
Reduction	73%	90%	43%	-189%	50%	34%

Table 13. INEX BEP agreement across structural levels.

	Level 1	Level 2	Level 3	Level 4	Level 5	All levels
General agreement	34%	25%	40%	34%	25%	38%
Unanimous agreement	6%	0%	2%	1%	0%	0%

Table 14. Proportion of BEPs selected by individual participants across structural level.

Participant	Level 1	Level 2	Level 3	Level 4	Level 5	All levels
1	19%	0%	39%	41%	50%	38%
2	13%	33%	38%	36%	50%	36%
3	44%	67%	40%	46%	0%	42%
4	63%	0%	44%	15%	0%	34%
Overall average	35%	25%	40%	35%	25%	38%

The most common level for both ROs and BEPs (Table 12) is level 3, which mostly corresponds to sections within articles. There are more ROs than BEPs overall and at each individual level except at level 4, which mostly corresponds to sections within articles. Reduction is comparatively high at higher structural levels and falls considerably at lower structural levels.

Both general and unanimous agreement (Table 13) are low. General agreement is highest at level 3, which mostly corresponds to sections of articles, and lowest at levels 2 (front matter, body, back matter) and 5 (e.g. sentence). Unanimous agreement is extremely low at all levels. Analysis of the BEPs selected by individual participants suggests that some participants had a preference for components at a particular structural level (Table 14). Furthermore, examination of participant logs shows that some participants actually had a preference for particular document components at the same level, e.g. abstract, introduction, conclusions. These findings may partly explain the low level of agreement between participants. The small-scale nature of the study may also have contributed to this result.

6. Conclusions and future work

This paper has introduced the concept of focussed retrieval, employing BEPs to provide optimal starting-points from which users can browse to find relevant objects. The work described here uses a document collection of Shakespeare plays marked up in XML, along with 43 queries, relevance assessments and BEPs gathered during the Shakespeare user study. Some preliminary findings from a small-scale experimental study using the INEX 2002 data have also been presented.

With regards to the general nature of BEPs, our results show that the concept of BEP is intuitive, as shown by the higher levels of agreement for BEPs than for ROs in the Shakespeare data, and the high reduction levels from ROs to BEPs across all query categories and structural levels. This conclusion is supported by the findings from the INEX data, which show medium to high reduction levels across almost all structural levels. It is also clear from our results that BEPs have a distinct character, as shown by the comparison between BEPs and (leaf and extrapolated) ROs. The nature of BEPs cannot simply be explained by extrapolated relevance; it is clear that additional factors are introduced by the structural nature of the documents and queries.

With regard to structural level of BEPs, the findings of the Shakespeare study, supported by the results from the further INEX study, indicate that BEPs are usually chosen at the same structural level as, or the level immediately above, the ROs. However, there were some indications in the INEX study that more specific BEPs (i.e. BEPs at a *lower* structural level) were sometimes chosen instead. This result may be partially due to the characteristics of the INEX relevance assessment sets. Since relevance assessors also judged ascendant components (parents) and descendant components (children) of the original relevant components, the resulting set of relevance assessments could contain several components, at different structural levels, that contained the same relevant information. For example, if a paragraph was judged to be relevant, the ascendant components (section, body, article) may also appear in the relevance assessment set for that topic. It was observed in (Ghuman, 2004) that, in such cases, participants very often chose a component at a lower structural level as a BEP. This is reflected in the corresponding BEP type analysis, where a high proportion (50%) of partRJ BEPs were identified.

With regard to agreement, BEP agreement was highest at lower structural levels in the Shakespeare study. This result was not reflected in the preliminary findings from the INEX study, where agreement was found to be highest at the middle structural level, and inconsistent at the remaining levels. This seemed to be partially due to the personal preferences of the participants involved in the study, who regularly chose components of the same type (e.g. abstract, introduction, conclusions), which belonged to this middle structural level. It is thus not clear if this result would extend to future studies using the INEX data.

By examining the effect of query category, we can see that essay-topic and content-only queries usually produce consistent results close to average values, both with regards to BEPs and their relationship with ROs, and agreement between participants. This is as expected, since these types of queries most closely approximate the standard, "topical" type of query used in information retrieval experiments. Factual queries, on the other hand, exhibit atypical behaviour, and content-and-structure queries exhibit inconsistent behaviour for many of the factors we examined.

In order to explore the practical application of our findings from this study, we developed a set of simple heuristics that could be used to identify BEPs from ROs. These heuristics were applied to the relevance assessments from the Shakespeare study, using different thresholds, in order to derive BEPs automatically (Lalmas & Reid, 2003); the derived BEPs were then compared with the BEPs obtained from the experimental participants during the study. Examples of the heuristics are:

- If the proportion of relevant lines in a container speech exceeds a given percentage threshold, the speech should be selected as a BEP.
- If relevant lines are evenly distributed throughout the container speech (determined by using a distribution formula and applying a threshold), the speech should be selected as a BEP.
- If the number of consecutive relevant lines exceeds a given threshold, the first line of the sequence should be selected as a BEP.
- If the number of consecutive non-relevant lines exceeds a given threshold, the first line after the non-relevant sequence should be selected as a BEP.

Analysis of the results of this experiment suggests that such simple rules are not sufficient, singly at least, to explain the identification of BEPs from ROs. The process of encoding human criteria for BEP selection is clearly very complex, and the different characteristics of individual query categories appear to provide a further complicating factor.

In parallel, (Finesilver and Reid, 2003) examined and compared the usage and effectiveness of BEPs and ROs for a single query category, namely essay-topic, content-only queries. One of the main findings of this study was that different queries belonging to the same query category elicited different information searching behaviour. Where an optimal strategy was employed, participants achieved better task performance using BEPs than ROs. From observation of the experiment and examination of the software logs, it appeared that failure to adopt an optimal strategy could be at least partially attributed to the inconsistent nature of the BEPs. Two different BEP types were identified: *browsing BEPs* (defined as "the first object in a sequence of relevant objects") and *container BEPs* (defined as "(real or virtual) objects which contained several relevant objects").

Taken together, the findings from all these studies suggest that further work is required to shed more light on the complex nature of BEPs and the process of deriving them from ROs. This work will focus on two main themes: (1) a more detailed examination of the Shakespeare data, in particular exploring the concept of BEP types; and (2) an empirical study of the usage and effectiveness of BEPs. This work is the focus of (Reid et al, 2005).

Acknowledgements

Section 5 of this paper builds on work carried out by Mandeep Kaur Ghuman (Ghuman, 2004) in the course of her BSc final year project in the Department of Computer Science, Queen Mary, University of London.

References

Brin, S., and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine, *7th WWW Conference*, Brisbane, Australia.

Cleveland, D.B., Cleveland, A.D., and Wise, O.B. (1984). Less than full text indexing using a non-Boolean searching model, *Journal of the American Society for Information Science ,* 35(1):19-28.

Chiaramella, Y., Mulhem, P., and Fourel, F. (1996). A model for multimedia information retrieval, Technical Report Fermi ESPRIT BRA 8134, University of Glasgow, 1996.

Craswell, N., Hawking, D., Wilkinson, R., and Wu, M. (2003). Overview of the TREC 2003 Web Track, *NIST Special Publication 500-255: The Twelfth Text REtrieval Conference (TREC 2003)*.

Finesilver, K., and Reid, J. (2003). User behaviour in the context of structured documents. *European Conference on Information Retrieval (ECIR2003)*, Pisa, Italy, pp 104-119.

Frisse, M. (1988) Searching for information in a hypertext medical handbook, *Communications of the ACM*, 31(7):880-886.

Fuhr, N., and Großjohann. K. (2001). XIRQL: a query language for information retrieval in XML documents, *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, USA, pp 172-180.

Fuhr, N., Malik, S., and Lalmas M. (2004). Overview of the INitiative for the Evaluation of XML Retrieval (INEX) 2003, *Proceedings of the Second INEX Workshop*. Dagstuhl, Germany.

Furnas, G.W. (1999). The fisheye view: A new look at structured files, in S.K. Card, J.D. Mackinlay & B. Shneiderman (eds.), *Readings in Information Visualization: Using Vision to Think*, pp 312-330.

Ghuman, M. (2004). *Creating and Identifying Different BEPs for INEX 2002*. BSc Final Year Project, Queen Mary, University of London.

Hertzum, M., and Frøkjær, E. (1996). Browsing and querying in online documentation: A study of user interfaces and the interaction process, *ACM Transactions on Computer-Human Interaction*, 3(2):136-161.

Hertzum, M., Lalmas, M., and Frøkjaer, E. (2001). How are searching and reading intertwined during retrieval from hierarchically structured documents? *INTERACT 2001*, Japan.

Kazai, G., Lalmas, M., Fuhr, N., and Gövert, N. (2004). A Report on the First Year of the INitiative for the Evaluation of XML Retrieval (INEX'02), *European Research Letter, Journal of the American Society for Information Science and Technology*, 55(6):551-556.

Kazai, G., Lalmas, M., and Reid, J. (2003). Construction of a test collection for the focussed retrieval of structured documents, *European Conference on Information Retrieval (ECIR2003)*, Pisa, Italy.

Kazai, G., Lalmas, M., and Rölleke, T. (2001). A Model for the Representation and Focussed Retrieval of Structured Documents based on Fuzzy Aggregation, *8th International Symposium on String Processing and Information Retrieval (SPIRE2001)*, pp 123-135, Laguna de San Rafael, Chile.

Kazai, G., Lalmas, M., and Rölleke, T. (2002). Focussed Structured Document Retrieval, *9th International Symposium on String Processing and Information Retrieval (SPIRE2002),* Lisbon, Portugal.

Kotsakis, E. (2002). Structured information retrieval in XML documents. *ACM Symposium on Applied Computing*, Special track on information access and retrieval, Madrid, Spain.

Lalmas, M., and Reid, J. (2003). Automatic Identification of Best Entry Points for Focussed Structured Document Retrieval, Poster, *CIKM Conference on Information and Knowledge Management*, New Orleans, Louisiana, USA, pp 540-543.

Myaeng, S., Jang, D.H., Kim M.S., and Zhoo Z.C. (1998). A flexible model for retrieval of SGML documents, *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, pp 138-145.

Pithia, J. (2002). *Best entry point algorithms for focussed structured document retrieval*. MSc Final Project in Advanced Methods of Computer Science. Queen Mary University of London.

Reid, J., Lalmas, M., Finesilver, K., Hertzum, M. Best Entry Points for Structured Document Retrieval – Part II: Types, Usage and Effectiveness. Information Processing & Management, 2005 (To appear).

Rölleke, T., Lalmas, M., Kazai, G., Ruthven, I., and Quicker, (2002). S. *The accessibility dimension for structured document retrieval, European Conference on Information Retrieval (ECIR2002)*, Glasgow, Scotland.

Rölleke, T. (1999). POOL: *Probabilistic Object-Oriented Logical representation and retrieval of complex objects - A model for hypermedia retrieval*, Ph.D. Thesis, University of Dortmund, Verlag-Shaker.

Srigananathan, G. (2002). *An investigation into methodologies for the automatic construction of test collections of XML structured documents*, BSc Final Year Project, Queen Mary University of London.

Tenopir, C., and Ro, J.S. (1990). *Full text databases*. Greenwood Press.

Vorhees, E.M. (1998). Variations in relevance judgments and the measurement of retrieval effectiveness, *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Melbourne, Australia, pp 315-323.

Wilkinson, R. (1994). Effective retrieval of structured documents, *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland, pp 311-317.