Information Processing & Management, vol. 42, no. 1 (2006), pp. 89-105.
Preprint version

Best Entry Points for Structured Document Retrieval – Part II: Types, Usage and Effectiveness

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. It investigates at the types of best entry points in structured document retrieval, and their usage and effectiveness in real information search tasks.

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. Structural information can be exploited at the presentation stage 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. In this paper, following from the work of (Chiaramella, Mulhem & Fourel, 1996) in the context of hypermedia multimedia retrieval, a best entry point (BEP) is defined as a document component from which the user can obtain optimal access, by browsing, to relevant document components. 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 (Reid et al, 2005), we described the methodology and results of a study of BEP characteristics that was carried out on the Shakespeare test collection (Kazai, Lalmas & Reid, 2003). This test collection consists 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. Results analysis focused on the relationship of BEPs to relevant objects (ROs) and agreement between experimental participants, and was performed across different query categories (factual/essay-topic; content-only/content-and-structure). Some preliminary findings from a small-scale experimental study using the INEX 2002 data were also presented.

This work showed that the concept of BEP is an intuitive one and that BEPs have a character distinct from that of ROs, introduced by the structural nature of the documents and queries. BEPs were often chosen at the same structural level as, or the level immediately above, the ROs, although this was not always the case in the INEX study. The results with regard to BEP agreement were inconsistent between the Shakespeare study and the INEX study. The results with regard to query

category showed that essay-topic and content-only queries usually produced consistent results close to average values, while factual queries exhibited atypical behaviour and content-and-structure queries exhibited inconsistent behaviour for many of the factors we examined.

In order to explore the practical application of our findings from this study, we then carried out a further study that involved the development of a set of simple heuristics that could be used to identify BEPs from ROs, and their application to the relevance assessments from the Shakespeare study in order to derive BEPs automatically (Lalmas & Reid, 2003). Analysis of the results of the 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 & 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 is the subject of this paper, and focuses 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** in real information search tasks.

The remainder of the paper is organised as follows. In Section 2, we describe our study of BEP types. We propose a categorisation of BEP types and examine the characteristics of these types across query categories and participants. In Section 3, we describe our study of BEP usage and effectiveness in the context of real information searching tasks, again analysed across query categories. In Section 4, we relate these findings to the small-scale user study we carried out using the INEX data set (Fuhr, Malik & Lalmas, 2004). We conclude the paper with a summary of our findings and proposals for future work in Section 5.

2. BEP Types

In (Finesilver & Reid, 2003), it was observed that there were at least two different types of BEPs: browsing BEPs, which were defined as the first object in a sequence of relevant objects, and container BEPs, which were defined as (real or virtual) objects that contain several relevant objects. In this section, we propose a full categorisation of BEP types (Section 2.1) and examine the characteristics of these types across query categories and participants (Section 2.2). Our overall findings are discussed in Section 2.3.

2.1. BEP Type Definitions

Categorisation of BEP types was performed retrospectively, based on the following data:

- Experimental observations, participant interviews and analysis of software logs from the user experiment that led to construction of the Shakespeare test collection.
- Experimental observations and analysis of software logs from the user experiment described in (Finesilver & Reid, 2003).
- Common-sense, intuitive analysis of the Shakespeare test collection data, i.e. attempting to identify patterns in the relationship of BEPs to ROs, using such factors as structural level and sequences of consecutive relevant and non-relevant objects.

Five types of BEPs were identified and can be defined as follows:

- ***Relevance judgement* (RJ):** An RJ BEP is identical to a relevance judgement for that query. Therefore, it is, by definition, a leaf-level BEP.

- ***Browsing* (Br):** A browsing BEP is the first relevant object in a sequence of relevant objects intended to be accessed from that BEP. Therefore, it is, by definition, a leaf-level BEP. There may be non-relevant objects within the sequence; however, the non-relevant objects should not span more than a fixed threshold. A reasonable threshold in this instance was determined to be two speeches.
- ***Container* (Cntnr):** A container BEP is a higher-level object that contains at least one relevant object.
- ***Combination* (Comb):** A combination BEP combines the characteristics of browsing and container BEPs. It is a higher-level object that contains at least one relevant object, and is the first of a sequence of similar objects at the same structural level.
- ***Context* (Cntxt):** A context BEP is a non-relevant object at any structural level that is intended to provide contextual information for a relevant object, or series of relevant objects, that follows it.

2.2. Results and analysis

In our analysis, we examine the distribution of BEP types and, for each BEP, the number of leaf-level objects that are accessible, the percentage of the total number of accessible objects composed of relevant objects, and the percentage of the total number of relevant objects composed of accessible objects. Analysis was performed across query categories using (non-merged) unique BEPs and across participants using non-unique BEPs, and average values were calculated accordingly.

Firstly, we examined the distribution of BEP types across query categories (Table 1). Overall, there was a similar proportion of browsing and container BEPs, and very few context BEPs. For factual queries, the most common BEP type was browsing, which may be due to the fact that such queries are usually answered by a small, often consecutive, number of leaf-level objects. For content-and-structure queries, container BEPs were the most common type, which may be due to the addition of a structural constraint, often explicitly indicating a higher-level object.

Table 1. Distribution of BEP types, by query category.

Query category	RJ	Br	Cntnr	Comb	Cntxt	All BEPs
Factual	1.27 (3%)	3.36 (8%)	1.55 (4%)	1.36 (3%)	0.45 (1%)	8.00 (18%)
Essay-topic	2.00 (4%)	4.13 (10%)	4.34 (10%)	3.25 (7%)	0.66 (1%)	14.38 (33%)
Content-only	2.00 (4%)	4.46 (10%)	3.77 (8%)	3.03 (7%)	0.63 (1%)	13.89 (31%)
Content-and-structure	1.00 (2%)	1.63 (4%)	3.00 (7%)	1.63 (4%)	0.50 (1%)	7.75 (18%)
Total	6.27 (14%)	13.57 (32%)	12.66 (29%)	9.27 (21%)	2.24 (5%)	44.01 (100%)

We also examined the distribution of BEP types for individual participants (Table 2). The overall pattern of distribution of BEP types is very similar for non-unique BEPs: browsing and container BEPs are the most common, and context BEPs the least common. The percentage of BEP types varies greatly between individual participants. It is also clear that individual participants often strongly favoured *either* leaf (RJ and browsing) *or* non-leaf (container and combination) BEPs. At either extreme, participant 1 chose 56% leaf-level BEPs and 40% higher-level BEPs, while participant 5 chose 23% leaf-level BEPs and 72% higher-level BEPs. Both RJ and context BEPs showed a high variability among participants.

Table 2. Distribution of BEP types for individual participants.

Participant	RJ	Br	Cntnr	Comb	Cntxt	All BEPs
1	2.00 (22%)	2.90 (34%)	1.60 (18%)	1.90(22%)	0.30 (3%)	8.70 (100%)
2	0.10 (2%)	1.50 (25%)	2.40 (39%)	1.50 (25%)	0.60 (10%)	6.10 (100%)
3	0.10 (1%)	1.70 (25%)	1.80 (26%)	2.60 (38%)	0.60 (9%)	6.80 (100%)
4	1.30 (12%)	4.20 (40%)	2.30 (22%)	2.60 (25%)	0.10 (1%)	10.50 (100%)
5	0.30 (4%)	1.40 (19%)	2.80 (37%)	2.60 (35%)	0.40 (5%)	7.50 (100%)
6	1.80 (20%)	2.70 (33%)	2.70 (31%)	1.40 (16%)	0 (0%)	8.60 (100%)
7	0.36 (5%)	3.18 (41%)	2.09 (27%)	1.55 (20%)	0.55 (7%)	7.73 (100%)
8	0.55 (9%)	1.09 (19%)	2.18 (38%)	1.91 (33%)	0.09 (2%)	5.82 (100%)
9	2.00 (19%)	2.18 (22%)	3.91 (36%)	2.45 (22%)	0.18 (2%)	10.73 (100%)
10	2.17 (23%)	2.25 (25%)	3.00 (31%)	1.83 (19%)	0.25 (3%)	9.50 (100%)
11	0.55 (9%)	2.00 (34%)	1.73 (30%)	1.09 (19%)	0.45 (8%)	5.82 (100%)
Overall average	1.02 (11%)	4.56 (29%)	2.41 (30%)	1.95 (2%)5	0.32 (4%)	7.98 (100%)

To examine variation between participants more fully, we calculated the proportion of unique BEPs of each type chosen by each participant (Table 3). Overall, participant 3 chose the smallest percentage of BEPs (26%), while participant 7 chose the greatest percentage of BEPs (46%). RJ, browsing and contextual BEPs all show a high degree of variation between participants, while combination BEPs show the least variation, showing that all participants could identify a good proportion of cases in which combination BEPs were useful.

Table 3. Proportion of unique BEPs of each type chosen by individual participants.

Participant	RJ	Br	Cntnr	Comb	Cntxt	All BEPs
1	40%	34%	27%	33%	33%	33%
2	14%	32%	33%	23%	60%	30%
3	14%	22%	23%	34%	27%	26%
4	48%	49%	27%	30%	11%	36%
5	27%	26%	42%	41%	57%	37%
6	40%	43%	39%	35%	0%	39%
7	57%	56%	48%	35%	35%	46%
8	30%	19%	34%	32%	17%	28%
9	42%	41%	44%	45%	33%	43%
10	49%	44%	36%	38%	50%	41%
11	29%	24%	31%	27%	33%	28%
Overall average	36%	36%	35%	34%	32%	35%

Secondly, we examined the number of leaf-level objects accessible from individual BEPs, and averaged these figures across query categories and BEP types (Table 4). Since RJ BEPs consist of a single relevant object, they always support access to exactly one leaf-level object. Overall, averaged across queries, combination BEPs supported access to the most relevant objects, followed by context and container BEPs, followed by browsing and RJ BEPs. It is also clear from these results that BEPs have different characteristics in the context of different query categories: factual and content-and-structure queries seem to share similar characteristics, and essay-topic and content-only queries also seem to be similar.

Table 4. Number of leaf-level objects accessible from BEPs, by query category and BEP type.

Query category	RJ	Br	Cntnr	Comb	Cntxt	All BEPs
Factual	1.00	26.83	12.43	16.58	43.67	12.77
Essay-topic	1.00	6.81	20.09	33.69	18.47	50.13
Content-only	1.00	6.93	21.88	29.91	18.49	45.14
Content-and-structure	1.00	32.42	10.66	34.72	37.89	27.12
Overall average	1.00	7.33	20.38	31.30	22.07	43.44

Table 5. Number of leaf-level objects accessible from BEPs, by participant and BEP type.

Participant	RJ	Br	Cntnr	Comb	Cntxt	All BEPs
1	1.00	6.18	42.74	18.98	4.25	40.27
2	1.00	4.78	16.12	55.00	37.11	60.98
3	1.00	9.38	19.90	23.74	15.58	38.86
4	1.00	12.50	35.46	29.78	4.00	51.49
5	1.00	8.43	25.70	44.03	24.33	66.62
6	1.00	11.01	13.86	60.31	-	61.71
7	1.00	6.39	12.44	32.77	25.63	45.99
8	1.00	13.86	33.72	32.25	13.50	52.39
9	1.00	4.60	7.01	30.29	19.00	35.06
10	1.00	5.71	7.76	35.59	13.00	35.36
11	1.00	6.64	47.61	24.57	19.38	53.24
Overall average	1.00	8.13	23.85	35.21	17.58	49.27

We also averaged across participants (Table 5). Participant 6 did not choose any context BEPs for any of his/her queries. It is clear that there is considerable variation among participants: although it is possible to discern that some pairs of participants

share similar profiles, e.g. participants 5 and 8, and participants 9 and 10, there is no obvious common pattern of BEP characteristics across all participants, or even across groups of participants.

Thirdly, we examined the proportion of *relevant* accessible objects, i.e. the proportion of objects accessible from individual BEPs that is composed of *relevant* objects (Table 6). For RJ BEPs, this figure is 100%, by definition. Among the other BEP types the figure is highest for browsing BEPs (93%), and lowest for context BEPs (53%). This could be expected, since browsing BEPs most often start solid sequences of relevant lines, while context BEPs are themselves non-relevant objects, and may be followed by further non-relevant objects before any relevant objects are encountered. Across query categories, the proportions are rather similar. As seen for the total number of accessible leaf-level objects, described earlier in this section, factual and content-and-structure queries exhibit similar behaviour, while essay-topic and content-only queries also display similar characteristics.

Table 6. Proportion of *relevant* accessible objects, by query category and BEP type.

Query category	RJ	Br	Cntnr	Comb	Cntxt	All BEPs
Factual	100%	98%	68%	65%	26%	90%
Essay-topic	100%	91%	74%	71%	57%	81%
Content-only	100%	92%	75%	67%	57%	83%
Content-and-structure	100%	98%	67%	83%	25%	83%
Overall average	100%	93%	73%	70%	53%	83%

We also examined the proportion of relevant accessible objects across participants (Table 7), in order to assess whether participants used BEP types in differing ways. The figures for each participant were calculated by averaging across all queries performed by that participant. Overall, participants identified BEPs that covered a large proportion of relevant objects: participant 10 chose BEPs with the maximum coverage (92% of relevant objects), and participants 1 and 3 chose BEPs with minimum coverage (75% of relevant objects). Context BEPs show the greatest variation among participants, with participant 1 choosing context BEPs covering only 5% of relevant objects (participant 6 did not identify any context BEPs at all), while participants 10 and 8 chose BEPs covering 89% and 86% of relevant objects, respectively. The least variation was observed for browsing BEPs, with participant 2 choosing BEPs with the maximum coverage (100% of relevant objects), and participant 10 choosing BEPs with minimum coverage (87% of relevant objects).

Table 7. Proportion of *relevant* accessible objects, by participant and BEP type.

Participant	RJ	Br	Cntnr	Comb	Cntxt	All BEPs
1	100%	96%	69%	64%	5%	75%
2	100%	100%	64%	73%	38%	76%
3	100%	93%	59%	70%	48%	75%
4	100%	96%	85%	85%	29%	89%
5	100%	96%	77%	85%	63%	86%
6	100%	89%	90%	76%	0%	91%
7	100%	94%	74%	82%	56%	85%
8	100%	90%	82%	62%	86%	79%
9	100%	90%	80%	76%	62%	88%
10	100%	87%	91%	76%	89%	92%
11	100%	91%	68%	68%	32%	77%
Overall average	100%	93%	76%	74%	51%	83%

Finally, we examined the proportion of relevant *accessible* objects, i.e. the proportion of relevant objects for a given query that is accessible from individual BEPs. This was found to be 100% for all queries: this means that, for each query, at least one participant chose a BEP that covered each of the relevant objects. The analysis across participants and BEP types is shown below (Table 8).

As could be anticipated from the BEP type definitions, the highest proportion of relevant objects is accessed through BEPs from a higher structural level, namely combination BEPs and container BEPs, while the lowest proportion of relevant objects is accessed through RJ BEPs (i.e. single leaf level objects). Ten of the eleven participants chose BEPs that supported access

to at least 95% of all relevant objects for their queries. This indicates that most participants thought the choice of relevant objects by other participants quite reasonable, even if it did not match with their own, and were willing to accommodate this in their choice of BEPs. Again, although no overall pattern can be established across all participants, it can be seen that some participants shared a similar strategy, e.g. participants 9 and 10.

Table 8. Proportion of relevant *accessible* objects, by participant and BEP type.

Participant	RJ	Br	Cntnr	Comb	Cntxt	All BEPs
1	17%	35%	47%	48%	6%	96%
2	20%	24%	55%	57%	21%	99%
3	6%	36%	36%	45%	16%	95%
4	9%	35%	38%	42%	5%	98%
5	2%	12%	45%	61%	34%	99%
6	9%	31%	47%	47%	0%	99%
7	3%	38%	27%	55%	39%	100%
8	9%	29%	55%	55%	40%	96%
9	9%	26%	30%	57%	40%	97%
10	17%	35%	26%	63%	29%	89%
11	15%	39%	59%	35%	31%	96%
Overall average	10%	31%	42%	51%	26%	97%

2.3. Discussion

Regarding the distribution of BEP types, the most common BEP types identified in the Shakespeare data were browsing and container BEPs. However, the percentage of BEP types varied greatly between individual participants. Individual participants also often appeared to have a preference for either leaf (RJ and browsing) or non-leaf (container and combination) BEPs.

With regards to the number of leaf-level objects accessible from individual BEPs, combination BEPs supported access to the greatest number overall. However, in this particular respect, there was no discernable pattern of BEP characteristics across all participants, or even across groups of participants. In other words, although participants seem to have an intuitive understanding of a BEP as a concept separate and distinct from ROs, there does not appear to be any shared understanding of the nature of different BEP types.

Overall, browsing BEPs provided access to the highest proportion of *relevant* accessible objects, and context BEPs the lowest proportion. Context BEPs also showed the greatest variability between participants in this respect. Combination and container BEPs provided access to the highest proportion of the total number of relevant objects, and RJ BEPs the lowest proportion, as was expected. No common pattern was discernable among participants in this respect.

The BEP type definitions employed in this analysis were derived from retrospective examination of the Shakespeare collection, and may therefore have been heavily influenced by the subject domain and structure of the data. Subsequently, in order to examine this issue, a further small-scale study was performed using a sub-set of the INEX 2002 test collection. The findings of this preliminary study, relating in particular to BEP types, are reported and discussed in Section 4.

It is also clear from the above analysis that BEPs have different characteristics in the context of different query categories. This finding supports the conclusions from our previous study that the influence of query category is very strong - see (Reid et al, 2005). Specifically, essay-topic and content-only queries share similar, "typical" characteristics, while factual and content-and-structure queries exhibit different, "atypical" behaviour. In fact, this factor seems a much more consistent determinant of participant behaviour than BEP types.

3. BEP usage

Our final study investigates the usage and effectiveness of BEPs in the context of real information searching tasks. It builds on a previous user study, described in (Finesilver & Reid, 2003), which examined the usage and effectiveness of BEPs compared to ROs, using only essay-topic, content-only queries from the Shakespeare study. The study described below uses queries belonging to each of the query categories (including one of the essay-topic, content-only queries used in (Finesilver & Reid, 2003), and focuses on comparing the usage and effectiveness of BEPs for these different categories. The

experimental methodology was designed to replicate, as closely as possible, the methodology used in the previous study, so that results could be compared across the two studies where appropriate. The following sections describe the interface used to display the BEPs (Section 3.1), the experimental methodology employed (Section 3.2) and the results and analysis (Section 3.3). Section 3.4 discusses the findings in the context of previous, related work.

3.1. Interface Development

In (Finesilver & Reid, 2003), an interface was developed in order to compare the usage and effectiveness of BEPs and ROs for two essay-topic, content-only queries. The interface for this second user study was constructed to replicate this design, so that any differences in results could not be attributed to the use of a different interface. The interface was designed to support both linear (at the same structural level) and hierarchical (moving between structural levels) information searching behaviour in the context of structured documents. The interface is split into two main panes. The left-hand pane displays the structure of the play. Any object that is opened up by the user is displayed in the right-hand pane, e.g. if the user clicks on a speech in the left-hand pane, the contents of that speech are displayed in the right-hand pane. Direct hierarchical information searching behaviour is supported by use of the left-hand structure pane. Linear information searching is supported by use of the previous and next buttons in the top right corner of the interface, or by linear use of the left-hand structure pane (e.g. using the keyboard arrow keys).

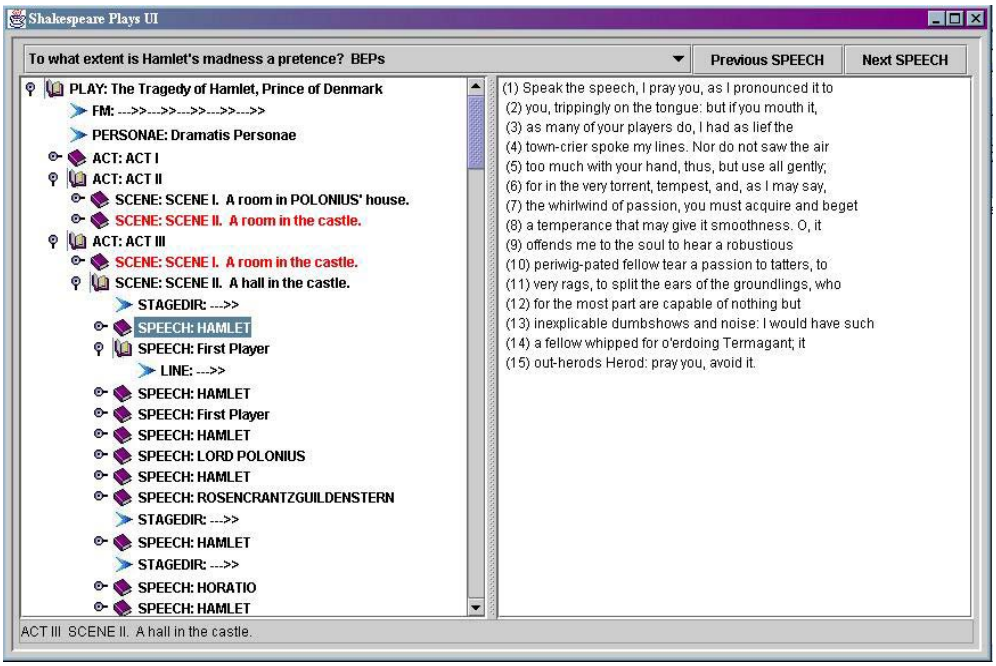

Figure 1. Search interface from (Finesilver & Reid, 2003).

Queries were pre-entered and chosen from a pull-down menu at the top of the screen. Software logging was set up, in order to log, for each object selected by the participant, the object ID, the time at which it was selected, its structural level and whether the object was a BEP. A separate log was created for each participant session.

3.2. Experimental Methodology

In this section, we discuss the queries, participants and experimental design. Four queries were chosen from the Shakespeare test collection, one query of each possible query category (i.e. query complexity / query type combination). Since the aim of the experiment was to encourage and examine information searching behaviour, the queries were chosen to be complex and have a high number of BEPs. The queries selected were:

- *Query 1*: "Give an analysis of the 'merry war' and witticism that pass between Beatrice and Benedick." (Essay-topic / content-only query relating to the play Much Ado About Nothing). This query was also used by Finesilver and Reid (2003).

- *Query 2*: "What might be the symbolic implications of the end of the play when Fortinbras is left the ruler of Denmark?" (Essay-topic / content-and-structure query relating to the play Hamlet).
- *Query 3*: "At what point does Macbeth decide to murder King Duncan?" (Factual / content-only query relating to the play Macbeth).
- *Query 4*: "In the beginning of the play A Midsummer Night's Dream, who was in love with whom? How did this change in the end?" (Factual / content-and-structure query relating to the play A Midsummer Night's Dream).

Table 9 shows a breakdown of the BEPs for these four queries, organized by structural level.

Table 9. Number of BEPs in the chosen queries, by query category and structural level.

Play (Query category)	Line	Other leaf	Speech	Scene	All BEPs
Much Ado (E/C)	10	0	22	3	35
Hamlet (E/CS)	2	1	5	1	9
Macbeth (F/C)	11	0	4	0	15
Midsummer (F/CS)	6	4	13	1	24
Overall average	7.25	1.25	11	1.25	20.75

Eight undergraduate students were recruited from the Department of English and Drama at Queen Mary, University of London. A Masters student was also recruited from the same department to mark the final task outcomes.

The experiment adopted a within-subject design, i.e. each participant performed all four queries. The order of the queries was counterbalanced in order to avoid any possible order effect. At the start of the session, each participant was given the chance to become familiar with the interface during a 10-minute training session, using a different Shakespeare play: during this time, the experimenter answered any questions relating to the interface functionality. The participant was then asked to fill in a background questionnaire, prior to starting on the query tasks. The background questionnaire elicited personal information (e.g. gender and year of study) and play-related information (general familiarity with the plays of Shakespeare, and familiarity with the individual plays to which the queries related).

The participant performed each query task in the given order. Performing a task involved: using the information searching interface to find text to answer the query; writing up a brief answer to the query; and completing a task questionnaire, which elicited information such as whether the participant had a prior idea of the answer to the query, how easy they found the query to understand, how difficult they found the query to answer, how useful they found the BEPs, and whether they had enough time to complete the query. There was no time limit, but the length of time taken by each participant was logged, as were the individual actions taken by the participant during each query (i.e. which objects they clicked on).

Finally, the participant was asked to fill in an interface questionnaire, which elicited their opinions about the interface and also gave them the opportunity to provide any other comments. After the experiment, the participants' answers were given to the marker, who assigned a score out of 25 to each answer, on the basis of quality and coverage.

3.3. Results and analysis

There are 3 main sets of results: data from the participant questionnaires, analysis of the BEPs chosen by participants during searching, and task performance data in the form of times and scores. Where possible, results were tested for significance using a two-way ANOVA (related) test. This allows us to examine the effect of query complexity, query type, any interaction effect between these two variables, and any participant effect (i.e. unwanted variations between participants). Only those results that relate directly to the issues of the usage and effectiveness of BEPs are described in detail below.

3.3.1 Questionnaire data

The data elicited by the participant questionnaires was divided into three types: background, task-related and interface-related. The background questionnaire elicited personal and play-related information. Four female and four male participants took part in the experiment. All the participants were in the final year of an undergraduate degree in English and Drama. The remaining background information concerns the participants' general familiarity with the works of Shakespeare, and their

familiarity with the individual plays used in the experiment (on a scale of 1 to 5, where 1 = very familiar and 5 = not familiar at all).

The average general familiarity of the participants with the works of Shakespeare was 2.125. The figures for familiarity with individual plays are shown in Table 10. Query complexity was found to be significant at $p<0.05$, i.e. participants were significantly more familiar with the plays for which they performed essay-topic queries than with those for which they performed factual queries.

Table 10. Participants' familiarity with individual plays.

Play (Query category)	Familiarity
Much Ado (E/C)	3.875
Hamlet (E/CS)	2.500
Macbeth (F/C)	2.375
Midsummer (F/CS)	2.750
Overall average	2.875

The task-related questionnaire examined whether the participant had a prior idea of the answer to the query (Table 11), how easy they found the query to understand (Table 12), how difficult they found the query to answer (Table 13), and how useful they found the BEPs (Table 14). The significant results were:

- More participants had a prior idea of the answer for the content-and-structure queries than for the content-only queries ($p<0.01$).
- There was variation among the participants with respect to ease of understanding the query, i.e. some participants found it easier to understand the set of queries, as a whole, than others ($p<0.05$).
- There was variation among the participants with respect to usefulness of BEPs, i.e. some participants considered the BEPs, overall, to be more useful than other participants did ($p<0.025$).

The interface-related questionnaire examined whether the participants thought the interface was straightforward, and how easy they found the interface to use. This information was gathered in order to verify that poor interface design had not influenced the results. None of the participants thought the interface was difficult to use or experienced any significant problems in using it.

Table 11. Number of participants with a prior idea of the answer to the query.

Play (Query category)	Prior idea
Much Ado (E/C)	3
Hamlet (E/CS)	6
Macbeth (F/C)	5
Midsummer (F/CS)	8
Overall average	5.5

Table 12. Ease of understanding query (1 = very easy, 5 = very difficult).

Play (Query category)	Ease of understanding
Much Ado (E/C)	2.750
Hamlet (E/CS)	2.500
Macbeth (F/C)	2.000
Midsummer (F/CS)	2.500
Overall average	2.438

Table 13. Difficulty of answering query (1 = very difficult, 5 = very easy).

Play (Query category)	Difficulty of answering
Much Ado (E/C)	2.750
Hamlet (E/CS)	3.000
Macbeth (F/C)	3.000
Midsummer (F/CS)	2.625
Overall average	2.844

Table 14. Usefulness of BEPs (1 = very useful, 5 = not at all useful).

Play (Query category)	Usefulness
Much Ado (E/C)	2.500
Hamlet (E/CS)	2.250
Macbeth (F/C)	2.125
Midsummer (F/CS)	2.625
Average	2.375

3.3.2 BEP data

The BEPs examined by participants during searching were logged and analysed by query category and structural level, and the proportion of the total number of BEPs was calculated (Table 15). The significant results are:

- Participants examined a greater proportion of BEPs for essay-topic queries than for factual queries ($p<0.01$).
- With regard to essay-topic queries, participants examined a greater proportion of BEPs for those queries that were content-only than for those that were content-and-structure (interaction effect , $p<0.025$)
- With regard to factual queries, participants examined a greater proportion of BEPs for those queries that were content-and-structure than for those that were content-only (interaction effect, $p<0.025$)

Table 15. BEPs examined during searching, by query category and structural level.

Play (Query category)	Line	Other leaf	Speech	Scene	All BEPs
Much Ado (E/C)	5.13 (51%)	0 (0%)	15.75 (72%)	2.38 (79%)	23.25 (66%)
Hamlet (E/CS)	0.63 (31%)	0.38 (38%)	3.38 (68%)	0.88 (88%)	5.25 (58%)
Macbeth (F/C)	2.75 (25%)	0 (0%)	1.88 (47%)	0 (0%)	4.63 (31%)
Midsummer (F/CS)	0.63 (10%)	1.00 (25%)	8.50 (65%)	1.00 (100%)	11.13 (46%)
Overall average	2.28 (29%)	0.69 (31%)	7.38 (63%)	1.42 (89%)	11.06 (50%)

3.3.3 Task performance data

The time spent by each participant on each query task was recorded (Table 16), and the task outcomes marked out of 25, giving each participant a possible total of 100 (Table 17). The significant results are:

- Participants scored better on factual queries than essay-topic queries ($p<0.05$)
- Participants scored better on content-only queries than content-and-structure queries ($p<0.025$)

Table 16. Time spent on query tasks (seconds).

Play (Query category)	Time (seconds)
Much Ado (E/C)	762.75
Hamlet (E/CS)	834.75
Macbeth (F/C)	653.00
Midsummer (F/CS)	736.38
Overall average	746.72

Table 17. Marks for query tasks (out of 25).

Play (Query category)	Mark (out of 25)
Much Ado (E/C)	17.50
Hamlet (E/CS)	15.88
Macbeth (F/C)	20.88
Midsummer (F/CS)	17.25
Overall average	17.88

3.3.4 Discussion

Some general conclusions can be drawn from consideration of the results of this user study in conjunction with the findings of (Finesilver & Reid, 2003). The results of this study show that the proportion of BEPs examined is higher for essay-topic queries than for factual queries, despite the fact that participants were more familiar with essay-topic queries than factual queries. Across all four queries, participants examined, on average, 50% of BEPs, but this figure rises to 58% for the essay-topic / content-and-structure query (Hamlet) and 66% for the essay-topic, content-only query (Much Ado). These figures are both higher than the figures of 51% and 46% for the two essay-topic / content-only queries used in the previous study (the latter figure of 46% relates to the same Much Ado query used in this study). Overall, it is clear that users are not likely to examine all the highlighted BEPs for a given query, even if they have unlimited time to do so.

It also appears that the particular properties of factual and content-and-structure queries elicit different information searching behaviour from participants. With regards to essay-topic queries, participants in this study examined significantly more BEPs for those queries that were content-only than for those that were content-and-structure. Taken in conjunction with previous results, we can conclude that this is likely to be due to the restrictive effect of the structural constraint in content-and-structure queries, meaning that participants already have a good idea of where to look for the answer to the query. With regards to factual queries, on the other hand, participants in this study examined significantly more BEPs for those queries that were content-and-structure than for those that were content-only. Since factual queries can generally be answered by reference to a very small number of texts, the structural constraint in content-and-structure queries may actually *increase* the number of texts that require consideration.

In this study, participants obtained higher scores for factual queries than for essay-topic queries. This is probably due to the simple fact that factual queries are easier to answer, and that it is usually clear when the "correct" answer has been found. Participants obtained higher scores for content-only queries than for content-and-structure queries, despite the fact that they more often had a prior idea of where to find the answer to content-and-structure queries than content-only queries. Since more BEPs were examined for essay-topic / content-only queries than for essay-topic / content-and-structure queries, it may be that BEPs were of particular help to participants in this respect.

Overall, the above results suggest that BEPs may be most useful for queries where there are many relevant objects, which are distributed widely throughout the text, i.e. essay-topic, content-only queries. This conclusion is not directly supported by the participants' estimations of BEP usefulness in this study. However, it does accord with the findings of the previous study by (Finesilver & Reid, 2003), who found that the adoption of an *optimal* strategy for using BEPs with essay-topic / content-only queries could produce improvement in performance (in terms of participants' scores) over usage of relevant objects.

4 BEPS in INEX

In (Reid et al, 2005), we described a related small-scale experimental study using the INEX2002 test collection (Fuhr, Malik and Lalmas, 2004). This further study was carried out 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).

The participant data from this study was examined to establish if the BEP types identified in the Shakespeare analysis also applied in the context of the INEX 2002 data, and if any further BEP types could be identified (Table 18). This analysis shows that combination and container BEPs were hardly used by the participants. In fact, the majority (55%) of the BEPs chosen by the participants could not be classified under any of the existing BEP types. Two new BEP types identified were:

- PartRJ (partial RJ). This was a very common type, accounting for 50% of all BEPs. In such cases, participants chose a sub-part of a relevance judgement as a BEP. For example, a participant chose as a BEP a particular paragraph from a section that had been judged as an RJ. The preponderance of this BEP type illustrates the shift towards lower level structural components already observed from the increased number of BEPs compared to ROs at structural level 4.
- New. In such cases, there was no discernable relationship between the chosen BEP and the ROs, i.e. the BEP was related to the ROs neither by hierarchical structure (e.g. container or partRJ BEP) nor by sequential structure (e.g. browsing or context BEP). The BEP was, therefore, marked as "new".

Table 18. Distribution of BEP types.

RJ	Br	Cntnr	Comb	Cntxt	PartRJ	New	All BEPs
47 (20%)	48 (20%)	1 (0%)	0 (0%)	11 (5%)	119 (50%)	13 (5%)	239 (100%)

All the participants identified a very similar total number of BEPs, and the distribution of these across BEP types was also similar for all participants (Table 19). However, as demonstrated by the low levels of agreement between participants (Reid et al, 2005), this does not mean that participants chose the same BEPs.

Across all participants, the proportion of unique BEPs chosen was similar for most BEP types, with a slightly higher percentage of browsing BEPs chosen than other BEP types (Table 20). There was least variation between participants in the percentage chosen for browsing and partRJ BEPs (an 8% range), All the participants chose a similar total percentage of unique BEPs, within an 8% range. However, there is little evidence of common patterns of BEP type choice across participants, although it is interesting that the single container BEP was chosen by all participants, and that the two participants with the highest overall percentages of BEPs chosen identified a considerably lower percentage of context BEPs than the other two participants.

Table 19. Distribution of BEP types for individual participants.

Participant	RJ	Br	Cntnr	Comb	Cntxt	PartRJ	New	All BEPs
1	18 (20%)	23 (25%)	1 (1%)	0 (0%)	1 (1%)	44 (48%)	4 (4%)	91 (100%)
2	15 (18%)	20 (24%)	1 (1%)	0 (0%)	5 (6%)	41 (48%)	3 (4%)	85 (100%)
3	23 (23%)	24 (24%)	1 (1%)	0 (0%)	2 (2%)	44 (44%)	7 (7%)	101 (100%)
4	18 (22%)	20 (24%)	1 (1%)	0 (0%)	5 (6%)	34 (41%)	4 (5%)	82 (100%)
Overall average	18.5 (21%)	21.75 (24%)	1 (1%)	0 (0%)	3.25 (4%)	40.75 (45%)	4.5 (5%)	90 (100%)

Table 20. Proportion of unique BEPs of each type chosen by individual participants.

Participant	RJ	Br	Cntnr	Comb	Cntxt	PartRJ	New	All BEPs
1	38%	48%	100%	0%	9%	37%	31%	38%
2	32%	42%	100%	0%	45%	34%	23%	36%
3	49%	50%	100%	0%	18%	37%	54%	42%
4	38%	42%	100%	0%	45%	29%	31%	34%
Overall average	39%	46%	100%	0%	29%	34%	35%	38%

5 Conclusions and Future Work

This paper has revisited 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 BEP types, it is difficult to discern any significant patterns in our results. The most common BEP types identified in the Shakespeare data were browsing and container BEPs. However, in the INEX study, the majority of BEPs could not be classified under any of the five existing BEP types, leading to the creation of two new types. In the Shakespeare data, the percentage of BEP types varied considerably between individual participants, although there was some evidence that

participants often had a preference for either leaf or non-leaf BEPs. In the INEX study, on the other hand, the distribution of BEPs across BEP types was similar for most participants, although participants did not agree on which specific BEPs should be chosen. In fact, the concept of BEP type, as currently conceived, appears to be a poor determinant of participant behaviour. Query category is a much more consistent determinant of behaviour in our results.

With regards to BEP usage and effectiveness, it is clear that users are highly unlikely to examine all indicated BEPs for a query, even if they have unlimited time to do so. In our experiment, participants examined, on average, only 50% of BEPs, and the maximum percentage of BEPs examined for an individual query was 66%. Again, query category was found to have a strong effect on our results. For essay-topic queries, the presence of a structural constraint appears to reduce the number of BEPs that are examined; for factual queries, on the other hand, the presence of such a constraint appears to increase the number of BEPs that are examined.

Finally, there is some evidence in our participant performance results that BEPs may be most useful for queries where there are many relevant objects, which are distributed widely throughout the text, i.e. essay-topic, content-only queries. This conclusion is not directly supported by the participants' estimations of BEP usefulness in this study, but it does accord with the findings of the previous study by (Finesilver & Reid, 2003).

Overall, it is obvious from the findings discussed in this paper and in (Reid et al, 2005) that the process of identifying BEPs is not straightforward. There are several factors that affect this process, some of which we have examined during the studies described in this paper, and some of which can be identified from our results. The factors are:

- The data used, specifically the subject domain and logical structure of the data. For example, participants in the INEX study often showed a preference for specific components, e.g. the article's abstract, while such a strategy was not possible with the Shakespeare data, where there were no equivalent "named" components. The more detailed and complex structure of the INEX data thus seemed to support participants in quickly identifying components that might be useful.

- The query (or topic) category. Analysis of the Shakespeare study data showed that the nature and characteristics of BEPs were very strongly influenced by query category. Essay-topic and content-only queries usually produced consistent results close to average values, while factual queries often exhibited atypical behaviour, and content-and-structure queries exhibited inconsistent behaviour. Although comparison with the INEX data is not possible, since the INEX study used only one topic, this result was confirmed by the study of BEP usage described in Section 4 of this paper. Indeed, it appears from our results that query category is the primary determinant of BEP-related characteristics and behaviour.

- The nature of the relevance assessments. In both the Shakespeare study and the INEX initiative, participants were provided with the relevance assessments for their given queries / topics, and asked to identify BEPs. It could therefore be assumed that there would be a strong relationship between the relevance assessments and the BEPs themselves. For example, the Shakespeare study relevance assessments were all at leaf level (subsequent analysis used the concept of extrapolated relevance to examine relevant objects at different structural levels), while the INEX relevance assessments were at any structural level. Furthermore, the INEX relevance assessment sets often included components at different structural levels that contained the same relevant information (i.e. parents of other relevant components). The differing nature of the relevance assessments in these two studies, particularly in terms of the structural level of the relevance assessments, appears to have had a strong effect, both on the relationship between BEPs and relevant objects (e.g. reduction levels) and on the nature of the BEPs themselves (e.g. distribution of BEPs and BEP types across structural levels).

- Experimental participants. In all the studies described in this paper, there was considerable variation between participants. While a certain level of variation must be expected in any experiment, there was some limited indication of common patterns and behaviour between small groups of participants in the analysis of the data from the Shakespeare study and subsequent BEP usage study. This finding is supported by (Ghuman, 2004), who observed participants' behaviour and strategies during the process of choosing BEPs for the given INEX topic. She

grouped their strategies into two categories: topic-based strategies, which resulted from examination of the topic statement, and searching strategies, which resulted from examination of the content and structure of the data itself. Certain strategies were adopted by all the participants, e.g. examining keywords (topic-based) and examining relevance assessments (searching), while other strategies were shared by sub-sets of participants, e.g. reading the narrative section of the topic (topic-based) and reading the first sentence of a paragraph to judge whether it was worth reading the rest of it (searching).

Future work will focus on four main areas. Firstly, we will conduct a larger-scale examination of the INEX data as part of a special INEX track. In 2004, INEX started four new tracks, one of which is the interactive track, which is planned to run for at least 2 years. The two principal aims of this track are to investigate the behaviour of users when interacting with components of XML documents, and to investigate and develop approaches for XML retrieval that are effective in user-based environments. The former aim is currently being investigated in INEX 2004, where our objective is to identify trends regarding how users access relevant information, without explicitly considering the notion of BEPs. We believe that this exploratory study will give further and new insights into how the browsing and querying paradigms may be combined to provide optimal access to relevant information items, i.e. to focus retrieval to the appropriate level of granularity. In 2005, we are planning a large-scale investigation, in which we will explicitly investigate the usage of BEPs in the context of the INEX data, based on the findings of the exploratory study from 2004.

Secondly, our results have shown that, although the concept of BEP type is a valid and interesting one, our current categorisation of BEP types does not extend to different document collections. It will therefore be necessary to perform a more rigorous analysis of document collections from different subject domains and with different structural characteristics, in order to develop a more comprehensive and generally applicable classification. The modified categorisation must incorporate the concepts of both hierarchical and linear structure in a consistent way.

Thirdly, although our results indicate that BEPs should be most useful for content-only, essay-topic queries (since such queries generally have many relevant objects widely distributed throughout the text), this finding is not supported by the opinions of the participants in the BEP usage study. In fact, in that study, participants rated BEPs as most useful for the factual, content-and-structure query, with the essay-topic, content-only query coming second. It is noticeable, however, that all the participants had a prior idea of the answer to the factual, content-and-structure query: it is possible, therefore, that participants were using the BEPs for this query as a navigation device in order to home in on an area of the text that they already knew to be relevant. In contrast, less than half the participants had a prior idea of the answer to the essay-topic, content-only query, and it is therefore likely that, for this query, participants were more heavily dependent on BEPs to direct them towards relevant information and support their browsing behaviour. Further controlled studies of BEP usage will, therefore, be necessary in order to examine this issue in more detail.

Finally, the conclusions drawn from these and future, related studies will be used to inform the design of prototype interfaces to a focussed SDR system. A candidate interface, the RelevanceLinkBar interface, has already been developed and has undergone an initial evaluation (Reid & Dunlop, 2003, Gkaranatsi, 2001). This interface, which extends the standard ranked retrieval interface of the Google search engine by providing the user with a visualisation of links within each webpage, has already been shown to provide good support for browsing behaviour. Current work in progress is comparing the relative effectiveness of representing in-links and out-links in the Relevance LinkBar. Future work in this area will therefore focus on implementing BEPs for essay-topic, content-only queries using the RelevanceLinkBar interface.

Acknowledgements

Parts of this paper build on work carried out by Mandeep Kaur Ghuman (Guhman, 2004) and Bertha Moron (Moron, 2004) in the course of their BSc final year projects in the Department of Computer Science, Queen Mary, University of London.

References

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

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.

Fuhr, N., Maalik, 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.

Gkaranatsi, Z. (2001). *A User-Oriented Evaluation of the RelevanceLinkbar Interface*. MSc Dissertation, Queen Mary, University of London.

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

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.

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.

Moron, B. (2004). *A User-centred Evaluation of the Effect of BEPs and Query Type / Query Complexity Combinations*. BSc Final Year Project, Queen Mary, University of London.

Reid, J., and Dunlop, M. (2003). Evaluation of a Prototype Interface for Structured Document Retrieval, *Human-Computer-Interaction conference (HCI 2003)* Bath.

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