COMPARISON OF SEARCH AND INDEXES
A Case Study

The benefits of indexes over search alone have frequently been outlined. Although there are many differences between search and indexes, the overarching argument against relying on search alone can be summed up as follows: searching retrieves far too many terms, many of which are time wasters that do not lead to relevant discussions; at the same time, relevant discussions are often missed.

This makes intuitive sense, but there may be a tendency to think that the difference is relatively insignificant, so that although search is not ideal it is "good enough". This has led me to investigate whether the difference is really enough to have a major impact on a reader's ability to find the information they need.

I therefore decided to do comparisons on two documents that I had indexed, to see what the differences were between the two methods.

Economic Consequences of the Peace

The first book that I considered was *Economic Consequences of the Peace*, by John Maynard Keynes. This book was written in 1919 by Keynes, a participant in the Paris Peace Conference at the end of World War I.

I chose five topics to try to access using search, and compared this to access using the index:

- reparation
- debt
- coal
- trade
- transfer of land/territory from Germany

Reparation

As my starting point I will assume that readers are always using the best search term; in this case it is "reparation" singular (rather than the usual modern term, reparations[1]).

I was presented with a mind-boggling 143 results. In comparison, the index entry had 37 page references in total, which were organised into subheadings with an average of 3 to 5 page references each; clearly a much easier task for the reader.

If the reader did have the patience to sift through the 143 results, would most of them lead to the same discussions covered by the index entry? In other words, is it mainly a case of the word being used many times in each discussion?

To find out I investigated each of the 143 mentions of "reparation". I found that only 52 of them led to relevant discussions. A "miss rate" of approaching two thirds was higher than I was expecting, so I took note of what these unindexed references were – perhaps valuable references had in fact been left out of the index and were only accessible by search?

However I found that these other mentions were indeed unindexable. The vast majority were titles of groups and documents that included the word "reparation", especially the "Reparation Commission" (which has its own entry) and the "Reparation Treaty", both of which did not fully overlap with discussions about the reparations, but were often mentioned in discussions about other topics. I felt sure that readers following up these other search results looking for information about the reparations would be left disappointed.

I then looked at the search versus index comparison from the other direction: were all of the indexed discussions covered by the search results? Given the number of search results, you might expect very good coverage of the index entries.

However I was surprised at the lack of overlap. Of the 37 locators in the index, only 19 were covered by the search results. In nearly half of the discussions on reparations, this word was not mentioned at all. The reparations were referred to in a great many ways, including indemnity, damages, amount due, amount owed, and liability. Often, no synonym was used at all, the topic was simply understood.

Debt

When I performed the search for "debt", (including "debt" as part of longer words[2]), I obtained 67 results. This compares to an index entry with only three subheadings and 9 page references (maximum five per subheading).

Of the 67 search hits, only 24 led to relevant references. Again, I combed through the text around the highlighted search results to judge what types of things were in the search but not the index entry. Given the huge disparity between 67 hits and 9 locators, I was particularly keen to see whether there was anything that, in hindsight, I wished I had indexed. I approached the task with an aim to persuade myself to include more in the index entry. However, I did not find anything else that I felt would not lead to a disappointed or frustrated reader.

There was no single category of references that made up the bulk of the extraneous search results; instead there was a very wide range of disparate passing or irrelevant mentions. In case it is difficult to imagine a reference to debt which is not considered a relevant discussion of debt, I have pulled out just a few typical examples:

> – Discussion of "popular delusion" that Belgium suffered most losses; Keynes' opinion that "it will turn out, I believe, that taking account of casualties, loss of property, and burden of future debt, Belgium has made the least relative sacrifice...".
>
> –Disparaging discussion of British election and how this led to Lloyd George making harsher demands of Germany: "a vote for a Coalition candidate meant the Crucifixion of Anti-Christ and the assumption by Germany of the British National Debt."
>
> –Naming of amount of liquid assets available for reparation "ultimately available after meeting private debts, etc"
>
> – Listing of different estimates of securities held by Germans in a selection of foreign countries; mention of somebody's estimate of the holdings by German nationals in the Turkish External Debt.

When I looked at how well the search terms covered the relevant discussions, i.e. the locators, I found that 7 of the 9 locators were accessible from search results. One of the missing discussions was a particularly important one, but on balance this was relatively good coverage.

One thing I did notice however, was that in several of the longer discussions, the search term came quite late on in the passage. For example, the main discussion of inter-Allied debt was on pages 133-

138, but the first occurrence of a word containing "debt" was on the fourth page of this passage. The main discussion of a US loan to Europe was on pages 139-141, but the word debt is first mentioned at the bottom of the second page, eight paragraphs into the discussion. If a reader is accessing these passages using search, there is a question mark over whether they would take the time to work their way back to the beginning of the discussion, or even realise that they had missed much of it. It seems likely that at least some readers would just start from the highlighted search term and read on from there, missing much of the information.

Coal

This search resulted in 107 hits, including longer words containing "coal". For this topic however, exact search parameters were less important; a search on only "coal" still resulted in 97 hits.

The 107 search results compared to an index entry with a total of 13 references (about three per subheading). Fifty-six of the search terms led to significant references, more successful than the previous searches, although still leaving roughly half of search results as likely to be disappointing.

The search results not corresponding to index entries often highlighted references that were passing rather than wholly irrelevant. They were often mentions of issues discussed elsewhere. For example:

> –In a listing of Germany's prewar economic activities; coal is just one of a number of items/activities in the list. The overall paragraph is a re-cap of a discussion previously given in more detail, and the subject is also covered again in more detail later.
>
> – Passing references to "coal situation" in discussions about other issues or topics, such as iron ore production and railways. The coal situation is not expanded on in these sections as the writer assumes the reader is aware of it from the more detailed discussions elsewhere.
>
> –In a listing of types of damages that it may or may not be possible to calculate – "the global land values, furniture, machinery, coal-lines, transport system, etc".

On some occasions, the term was used in topics not directly related to coal at all, for example references to coal regions (such as a discussion of a possible plebiscite among inhabitants of the Saar coal basin).

All of the relevant references to coal were covered by the search results, that is, all of the 13 locators could be accessed by searching for the term "coal". This is undoubtedly due to the lack of synonyms for coal, and in fact the difficulty of covering this topic at all without using the word.

Trade

The word "trade" gave me 57 results, compared to 10 subheadings and a total of 27 locators. Thirty-one of the 57 search results corresponded with indexable entries, and of the 27 locators in the index, 19 contained the term "trade"; the remaining eight were not accessible by searching.

Transfer of land/territory from Germany

This important topic is not quite as easily summed up by one term. The index uses "land, loss of" with cross-references from "territory" and "borders". In the search, "land" turned up 18 times, but none of these were at all relevant to the topic. "Border" receives one hit, which is again totally unrelated to the topic. "Territory" obtains 55 results, only 9 of which are relevant to the discussion.

The index entry has five subheadings and a total of 12 locators. Only 3 of these locators could be accessed by the search. The remaining discussions would be hard indeed to access with any specific term, other than the names of the territories being transferred, which only a knowledgeable reader would be able to do.

Anglo-Saxon Chronicles

This is a record of history and events in England from about 495 to 1154, with year by year entries, many written contemporaneously.

For this rather complex document I have investigated two concepts: the kingdom/sub-kingdom of Mercia, and the kingdom of Wessex.

Mercia

A search for "Mercia", "Mercian", or "Mercians" brings up 135 results[3],

This can be contrasted to a total of 28 locators covering this topic, across 12 subheadings.

Of the 135 search hits, just 19 related to relevant references to the kingdom/sub-kingdom of Mercia.

The remaining occurrences of the term generally were in passages that did have a link to Mercia, but were not considered relevant information about Mercia as a whole, i.e. the kingdom/sub-kingdom. For example:

– Details of the lives of kings, not considered integral in the history of Mercia (admittedly a subjective decision). Examples include a king having an argument with a servant, getting married, or dying peacefully. Successions and deaths were included in the Mercia index entry when these were considered to have an impact on history of the kingdom. Of course, an indexer might choose to include every succession and death in the entry for the kingdom. In this index, given the timespan of the Chronicles (approximately 700 years), this would have led to very long and cluttered entries for the kingdoms. In any case, only some of the successions and deaths would be covered by using search of *kingdom* names, as often only the name of the king is mentioned.

– Local gossip or church events, such as an abbot being appointed within Mercia.

– More distant connections to Mercia, such as reference to the grandson of the king of Mercia, or a bishop who had previously lived in Mercia.

This does show that a particular strength of search is for very detailed research, where there is not that much information about the topic generally, and the reader really wants to track down and tease out every possible morsel. This can be readily imagined with a topic like the Anglo-Saxon kingdom of Mercia, where a researcher or writer might be able to use even peripheral information to help build some kind of picture, given the paucity of original sources.

Would search alone be enough, however, if the reader does want an exhaustive survey of information about Mercia? Again, I assessed whether the indexed references would all be accessible by somebody using search.

There are 28 locators covering Mercia. Most (23) are covered by the search term, obviously quite a high proportion compared to the other searches I have investigated. However, I did find that the references not covered were:

– Passages about the most famous kings, e.g. Penda and Offa. Their contemporary fame made it unnecessary for the writers to specify their kingdom. In contrast, the lesser known kings were cited with their titles.

– Passages during the time when Mercia was split apart by the Vikings, where the terminology was more obscure, using vague references to "Vikings" that would be difficult to relate to Mercia except by very knowledgeable readers, or using the name Five Boroughs/Five Towns, also unlikely to be known to many readers.

So, the passages not covered by the search were actually particularly important ones.

Wessex

A combination of "Wessex", "West-Saxons", and "West-Saxon" yielded a total of 78 hits[4].

This was the one case where I was surprised by how low the number of search results was. After all, Wessex was a very important kingdom for several hundred years of the Chronicles, and involved in a lot of the most important action. There are certainly a lot more passages involving Wessex than Mercia, which gets almost twice as many hits.

In fact, the importance and dominance of Wessex during that period of history is directly linked to the limited use of these terms in the text. As mentioned above, the text often refers to the name of the king without reference to his kingdom, when he is a major, well known figure. The king of Wessex is very often a major figure (Alfred, Edward the Elder, Athelstan, Egbert, etc). The writer does not repeatedly refer to their country, just as a modern writer or journalist would not constantly be reminding you that President Obama is President of the United States, but would often simply say "Obama", "the President", etc.

In addition, major events, conquests, or battles are often described without Wessex being named as such, because the writer considered it a given.

As a result, only 31 of the 63 locators for the Wessex entry are accessible by the previously mentioned search terms, and the references that are left out are particularly likely to be the more important ones.

Conclusion

Uses of Search

There are two types of information needs where I would expect search to be relatively useful.

- If an exhaustive search is required, where even information peripheral to the topic is useful to that particular reader, search can clearly offer up some additional mentions for the time-rich researcher to investigate. For example, a historian might want to look at every reference to Mercia in the Anglo-Saxon Chronicles in order to glean information or inspiration that might be meaningful to him or her. As the above investigation makes clear however, if the search is not used in conjunction with the index, many references will still be missed, and there is a tendency for the more important references to be missed if only search is used.

- Obscure and/or very specific references not considered indexable would best be addressed via search. For example, if somebody is looking in the Anglo-Saxon Chronicles for brief mentions of Abbot Cuthbald, or Citizen Osric, or other obscure figures insufficiently important to be indexed, search is the obvious answer.

It is seems likely however that specific references would become less accessible by search the more important or commonly mentioned they are. For example, although the obscure Citizen Osric is likely to only be referred to as such, references to King Alfred may use alternative terminology such as "the King", "the army" (which is often used to describe the Viking army as well, with only the context making clear which has been referred to), "Wessex", "Ethered and his brother", etc.

Limitations of search

A significant drawback of relying on search alone is the amount of time and patience required by the reader. Realistically, what proportion of readers will be willing to follow up even 30 or 40 search results, let alone 100 or 150? An index makes the task so much more usable; the reader is likely to only be presented with about two to five pages or page ranges to follow up for a particular topic.

Term /concept	**No. of hits to follow up**	**Typical No. of pages/page ranges to follow up in index (for specific subheading)**	*Total No. of locators in index*	*Total No. of sub-headings in index*
Economic Consequences				
Reparation	**143**	**3-5**	*37*	*10*
Debt	**67**	**1-5**	*9*	*3*
Coal	**107**	**2-3**	*13*	*5*
Trade	**57**	**3**	*27*	*10*
Transfer of territory from Germany	**55**	**3**	*12*	*6*
Anglo-Saxon Chronicles				
Mercia/Mercian(s)	**135**	**2-3**	*28*	*12*
Wessex/West-Saxon(s)	**78**	**5-6**	*63*	*12)*

Secondly, the reader who does take the time to sift through this mass of results will tend to find that they are not well rewarded in this effort, as the bulk of the hits will lead to disappointment.

Term /concept	**No. of hits**	**No. of hits leading to relevant discussion**
Economic Consequences		
Reparation	143	52
Debt	67	24
Coal	107	56
Trade	57	31
Transfer of territory from Germany	55	9
Anglo-Saxon Chronicles		
Mercia/Mercian(s)	135	19
Wessex/West-Saxon(s)	78	66

A third and arguably even more serious disadvantage to relying on search alone is that, even if the search is executed perfectly, it is likely to simply miss some of the relevant discussions that the reader is looking for. As already noted, in my case studies this was as much as half of relevant discussions.

Term /concept	**Total No. of relevant references (i.e. indexed passages)**	**No. of references MISSED by search terms**
Economic Consequences		
Reparation	37	18
Debt	9	2
Coal	13	0
Trade	27	8
Transfer of territory from Germany	12	9
Anglo-Saxon Chronicles		
Mercia/Mercian(s)	28	5
Wessex/West-Saxon(s)	63	32

Moreover, search appears to be prone to miss out topics or discussions which are particularly important, as these are more likely to be referred to by a range of different terms, or simply considered to be understood. Even where many of the references are covered by the search, the ones which are excluded may be those which are of greatest importance or interest.

When topics are left out or underrepresented by search for any of these reasons, not only will the contents of the book be less usable and useful than they could be, but also the book will be especially disadvantaged with *potential* readers who are investigating it to see whether it has topics of interest and is worth reading/buying.

It is also worth noting that longer discussions may be partially missed out if search is used, as sometimes the highlighted term comes quite late on in the discussion, and readers may not realise this or take the time to scan backwards and discover where the relevant passage begins.

Another issue with search is that it also relies on the reader being familiar enough with the different types of search to be able to make reasoned choices about what type of search to carry out; without this, he or she might miss significant proportion of potential search results. It may be that in time, if search rules and methods are consistent across publishers and devices, knowledge of search choices becomes a given among most readers, so this becomes less of an issue. However, even so, search also relies on the reader being familiar enough with the book or document itself to be able to select exactly the correct search term and format, and to be able to judge if the search results they are getting are reasonable for that particular book. Without this familiarity, again a sometimes large proportion of potential search results will be missed out.

Notes

1. One particular issue about this term is that nowadays the term more commonly used is "reparations". The text always refers to "reparation", considering it a collective noun. A modern reader would be more likely to search for "reparations". Obtaining zero hits would probably encourage them to investigate why this major topic seemingly is never mentioned, and would hopefully lead them to figuring out that they need to search on the singular version of the term. However, they might not figure this out, especially if they have not read the book but are simply dipping into it to see if it includes information on the topic and is worth investigating further.

2. I found that searching for "debt" only (whole word search) gave me 26 results, whereas including debt as part of longer words gave me 67 results (including "debts" 20 times and "indebtedness" 18 times). Again, it is important that the reader is familiar enough with the text to be able to judge whether the search results seem reasonable or not.

3. If the reader searches for "Mercia" only, they will receive a scant 42 results. The bulk of the mentions are "Mercian" or "Mercians"; obviously, if the reader does not think of this, they will miss about two thirds of the results.

4. "Wessex" as a search term yielded only 30 results. "West Saxons" yielded zero results. "West-Saxons" combined with "West-Saxon" yielded an additional 48 results, for a total of 78 hits if the searcher takes the time to figure out the correct combination of terms.